DOI: 10.35191/medsoft_2020_1_32_16_21

REPREZENTACE A INTERPRETACE VÝSLEDKŮ GENETICKÉHO VYŠETŘENÍ: NÁVRH SYSTÉMU PERSONÁLNÍ GENETICKÁ KARTA

Michal Huptych, Lenka Lhotská

Anotace

V tomto příspěvku bychom chtěli popsat prvky potřebné ke komplexní a v klinické praxi využitelné reprezentaci výsledků genetického vyšetření. Tuto reprezentaci využíváme v návrhu systému pracovně nazvaném Personální genetická karta, který je připravován v rámci spolupráce CIIRC ČVUT a firmy Mediware a.s.

Genetické informace jsou v posledních letech stále častěji uvažovány v medicínských procesech i v oblastech, které nejsou přímo napojeny na lékařskou péči, ale se zdravím úzce souvisí, jako je např. nutriční poradenství. Velkým tématem je správné využití genetických informací ve vhodné formě. Z dosavadních zkušeností vyplývá, že samotné genetické vyšetření je pouze začátkem a je nutné k němu připojit také informace z něho vyplývající, jako jsou například predispozice k chorobám, známé metabolizace léků či intolerance k různým látkám – jinými slovy klinickou interpretací a doporučeními, které by měly být nedílnou součástí komplexní reprezentace genetického vyšetření.

Pro ukládání informací genetické vyšetření existuje několik široce používaných databází, jako jsou např. HGNC, NCBI RefSeq, NCBI dbNSP, HGVS a další. Struktura reprezentace genetické informace by měla umožňovat provázání těchto kódování s odkazem na příslušný zdroj kódu a získává tak zároveň i informace a znalosti obsažené v těchto databázích. Jako vhodný slovník pro popis měření a výsledku genetické analýzy se ukazuje být řízený slovník LOINC® (*www.loinc.org*), který představuje ověřený způsob reprezentace klinických a laboratorních analýz a je využíván v mnoha zemích světa. Tento systém umožňuje velmi komplexní reprezentaci genetického vyšetření.

V oblasti interpretací genetických vyšetření existuje také již několik ucelených databází jako jsou např. PharmGKB, která se zaměřuje primárně na farmakogenetiku. Právě v oblasti farmakogenetiky je přínos využívání genetické analýzy nejvíce patrný a má vysoký potenciál. Proto jsou naše záměry namířeny hlavně do oblasti farmakogenetiky a jejího provázání s farmakokinetikou a farmakodynamikou za účelem co nejlepší administrace léků a minimalizace rizik.

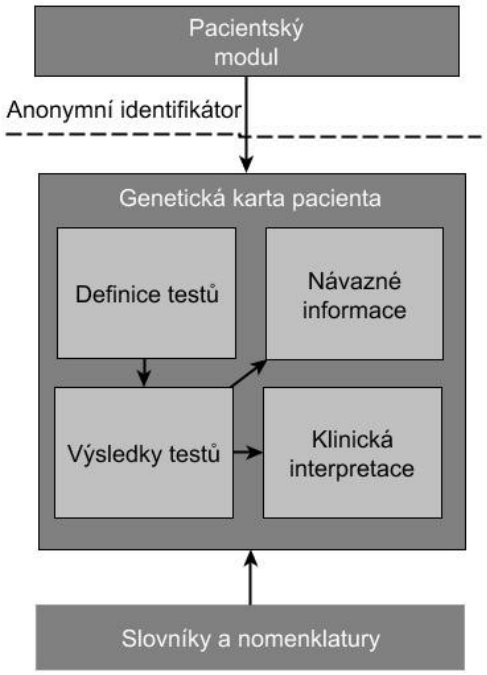
Klíčová slova

genetické vyšetření, interpretace genetického vyšetření, elektronický zdravotní záznam, standardy ve zdravotnictví

1 Úvod

Genetické vyšetření se v posledních letech celosvětově stává běžnějším procesem v rámci zdravotní péče i v oblastech jako je nutriční poradenství nebo sportovní medicína. Vzhledem k tomuto trendu je velkou otázkou, jakým způsobem je a jakým by mělo být zacházeno s těmito údaji. Genetické informace člověka jsou vesměs stálé, a tak má jednou provedené vyšetření víceméně celoživotní platnost a využitelnost. To, co činní genetické vyšetření důležitým z pohledu klinické medicíny, je vazba genetického profilu jedince na predispozice k různým chorobám nebo intolerancím nebo příslušná farmakogenetika. Avšak všechny tyto informace jsou založeny na empirických poznáních. V takovém případě je velmi nutné dbát na metodicky čistý a přesný výzkum, aby dosažené výsledky byly co nejvalidnější. A lze předpokládat, že se zvyšující se úrovní poznání bude docházet nejen k nárůstu propojení genetické informace a klinických dat, ale také k revizím a úpravám dřívějších závěrů. Z tohoto pohledu je vytvoření systému pro možné uložení genetické informace a její provázání na aktuální, validované interpretace logickým krokem. Avšak jsou tu dvě podstatné podmínky. Současně s nárůstem poznání propojení genetických a klinických informací, narůstá také míra využívání genetiky až téměř k jejímu zneužívání. Systém tedy musí mít jasně deklarovaný svůj záměr, který musí být jasně opodstatněný přínosem pro klinickou péči. V našem případě je takovým primárním záměrem zkvalitnění v oblasti administrace léků. Druhou podmínkou je nutnost velmi striktně a robustně chránit získaná data.

Základní schéma uspořádání navrhovaného systému nazvaného Personální genetická karta (PGK) je zobrazeno na obrázku 1.



Obrázek 1 – Přehledové schéma návrhu Personální genetické karty

Jak je z obrázku 1 patrné, je systém PGK je složen z 6 základních částí. Pacientská data by do PGK neměla vstupovat vůbec a uživatel je reprezentován pouze anonymním identifikátorem. Z druhé strany jednotlivé používané slovníky a nomenklatury jsou taktéž využívány, pokud možno přes definovaná API. Samozřejmě část těchto databází bude muset být součástí PGK. Důležitými částmi jsou 4 vnitřní moduly PGK. Pod tímto blokem je schována definice, co pro které klinické informace (např. pro účinnost určitých léků) má být z genetické informace testováno. Nad to je však v této části již jasně definována struktura a terminologie, která se má v rámci reprezentace a interpretace genetického testu využívat. V návaznosti na to jsou pak ukládány výsledky testů, jejichž reprezentace a struktura jsou jasně dány požadavky v definici testů. K těmto výsledkům jsou připojeny klinické interpretce a případné návazné informace. Které systémy slovníky a databáze mohou být využity při naplnění těchto 4 bloků je rozepsáno v následujících dvou kapitolách. Ve 4. kapitole se pak věnujeme souhrnnému pohledu na použití v PGK a v kapitole 5. se stručně zmíníme o možnost standardizované komunikace.

2 Reprezentace genetického vyšetření

V první řadě je nutné definovat způsob reprezentace genetického vyšetření. To lze rozdělit na 2 části. Jednak na definici struk-

tury reprezentace genetického vyšetření a terminologie, kterou chceme použít pro definici jednotlivých parametrů analýzy a jako definiční obor hodnot těchto parametrů.

Systémem, který jsme začali jako první zvažovat pro reprezentaci záznamu byl systém LOINC® (Logical Observation Identifiers Names and Codes) [1]. LOINC® je řízený slovník, přímo určený k vykazování výsledků měření a (laboratorních) testů, a je tedy vhodným kandidátem k použití v projektu. Druhým zvažovaným systém byla nomenklatura SNOMED CT® (Systematized Nomenclature of Medicine – Clinical Terms). Jeho cílové zaměření je však hlavně v oblasti interpretace na straně klinické praxe, je tedy vhodnější v oblasti definice hodnot. Vzhledem k tomuto faktu jsme se rozhodli zvolit jako základní reprezentaci systém LOINC®. Její stručný popis a příklad využití v genetické analýze je popsán v kapitole 2.1. Následně uvedeme základní přehled jednotlivých databází pro udávání hodnot jednotlivých parametrů genetického vyšetření s informacemi o propojení (provázání) údajů z jednotlivých databázích.

2.1 Základní struktura genetického vyšetření a systém LOINC®

Kódy, tabulky, poznámky a veškeré propriety řazené do systému LOINC [1] spravuje společnost Regenstrief Institute, Inc. a výbor Logical Observation Identifiers Names and Codes (LOINC) Committee, který má vyhrazena veškerá práva pro jeho správu a úpravy (kompletní znění podmínek použití je k dispozici na webu *https://loinc.org/terms-of-use*). K systému LOINC existuje nástroj Regenstrief LOINC Mapping Assistant (RELMA®), který je určen k vyhledávání v databázi LOINC. Stejně jako systém LOINC je i nástroj RELMA k dispozici zdarma za dodržení jistých podmínek (definováno společně s podmínkami LOINC v Terms of Use).

Každá z komponent v systému LOINC je definována unikátním kódem a příslušným názvem komponenty. Dále je v rámci komponenty definována řada informací ohledně její parametrizace a reprezentace případných výsledných hodnot. Pro parametrizaci je definována šestice typů údajů, jejichž kombinace udává specifikaci komponenty. Jedná se o následující parametry [1]:

- Component – tento kompletní název (Component) je přirozenou jazykovou reprezentací parametrizace daného konceptu
- Kind of Property (zkráceně Property) – uvádí, jakou veličinu (parametr) chceme v rámci daného konceptu udávat, resp. co bude výsledná hodnota konkrétně reprezentovat, např. Finding – nález; ID – identifikátor; Prid – přítomnost nebo identita; Num – číslo, Imp – interpretace
- Time Aspect (zkráceně Time) – definuje časový rozměr komponenty; základním časovým údajem (který je také používán v oblasti genetických testů) je Point time (Pt), tedy bod v čase
- System (Sample) – určuje na čem je analýza nebo měření prováděno; např. pro Sample Bld/Tiss (Blood or Tissue) je analýza prováděna z krve nebo tkáně
- Type of Scale (zkráceně Scale) – uvádí, jak je daná hodnota reprezentována; Qn – je to kvantifikátor, tedy číslo; Nom – hodnota je z nominálního seznamu, Ord – hodnota je z pořadového (ordinálního) seznamu, Nar – jedná se o volný text
- Type of Method (zkráceně Method) – udává, jaké metody bylo použito při analýze/testu; např. hodnota Molgen je zkratka pro Molecular Genetics

Ukažme si využití systému LOINC na příkladu jednoho ze základních panelů pro reprezentaci genetické analýzy [2] – komponentě 55233-1 Genetic Analysis Master Panel. Tento panel obsahuje tři části.

První je panel 55232-3 Genetic Analysis Summary Panel, který představuje shrnutí výsledků analýzy. Podle typu studie může souhrnný panel obsahovat proměnné, které shrnují farmakogenomickou studii nebo proměnné, které shrnují genetické nálezy, které jsou spojeny s chorobou nebo rizikem choroby.

Druhou částí je nepovinný panel výsledků – 55208-3 Genetic Analysis Discrete Result Panel, který obsahuje podrobný jednu a více komponent DNA 55207-5 Sequence Analysis Discrete Sequence Variation Panels. Tato poslední komponenta se může opakovat vícekrát tak, aby pokryl všechny zajímavé variace v rámci jednoho genetického vyšetření.

Komponenty, které reprezentují hodnotu v rámci testu, mají definován (doporučen) způsob, jakým výsledky reprezentovat. Velké množství výsledných hodnot je definováno jako položka z nominálního či pořadového seznamu. Pro tyto účely je vždy u dané komponenty určen slovník, nomenklatura či systém, v jejichž pojmech má být výsledek reprezentován. Samotný LOINC má definovánu skupinu tzv. odpovědí (Answers), které slouží jako číselníky pro výsledné hodnoty některých analýz. Velké množství výsledných hodnot je však doporučeno kódovat v rámci systémů a nomenklatur, které jsou popsány níže ve zbytku kapitoly 2. Případné získání těchto informací z příslušných nomenklatur je otázkou integrace datových zdrojů.

2.2 The HUGO Gene Nomenclature Committee (HGNC)

Jde o základní nomenklaturu pro kódování genů. Základní informace o HGNC jsou dostupné v [3]. Webový přístup je pře adresu *https://www.genenames.org,* přístup k databázi je dostupný na stránce *http://www.genenames.org/cgi-bin/statistics*. Data jsou vázána na jednotlivé chromozomy a jsou rozdělena do několika různých oblastí – protein-coding gene, non-coding RNA, pseudogene a other. Tyto údaje jsou dále děleny do dalších podskupin. Data je možné získat přes REST web-service, který poskytuje data ve formátu JSON nebo XML nebo stáhnout jako soubory ve formátu TXT nebo JSON nebo je možné přejít do grafického uživatelského rozhraní, kde lze přesně nastavit obsah, který má být v exportu. Vedle základních informací, jako je kód a název genu (např. HGNC:5 – alpha-1-B glycoprotein nebo HGNC:30005 – alpha 1,3-galactosyltransferase 2) mohou data obsahovat ještě množství dalších informací, např. synonyma nebo identifikátory NCBI RefSeq IDs, a další.

2.3 Reference Sequence Database (RefSeq NCBI)

Databáze NCBI RefSeq [4] je neredundantní, komplexní, komentovanou databázi, která obsahuje soubory sekvencí, včetně DNA transkripcí a bílkovin. Ještě než přejdeme k popisu RefSeq samotné, zmiňme v krátkosti NCBI (National Center for Biotechnology Information). Ta je provozovatelem a zaštítěním pro celou škálu různých projektů a databází, jako např. PubMed, Basic Local Alignment Search Tool (BLAST), nebo níže zmíněných dbSNP a ClinVar. RefSeq je dostupná na webu *https://www.ncbi.nlm.nih.gov/refseq*. Data databáze jsou dostupná v úložišti FTP *ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/*. Velmi důležitou součástí webu je pak stránka *http://www.ncbi.nlm.nih.gov/mapview/*, kde jsou odděleně uvedeny reference na různé živočichy. Pro potřeby projektu personální genetické karty by bylo potřeba mapování pro člověka. Stránky také obsahují grafícko-textový přehled pro jednotlivé chromozomy. Kódy z NCBI RefSeq jsou např. kód NG_007400.1, či NM_000088.3. Počáteční písmena kódu mají svůj význam. Definují totiž, jaké kategorie se daný gen týká. Toto rozdělení lze nalézt např. na *http://en.wikipedia.org/wiki/RefSeq*.

V rámci databáze RefSeq můžeme mít definováno několik kódů, které se týkají různých oblastí zároveň. Uveďme si toto na příkladu pro gen CYP1B1 (HGNC:2597), u kterého jsou v rámci databáze definovány následující kódy: NC_000002.12, NC_000002.11 – kódy pro kompletní genomické moleku-

ly (v podstatě říká, že gen se nachází na 2. chromozómu), NG_008386.1 – kód pro nekompletní genomické molekuly, NM_000104.3 – kód mRNA, NP_000095.2 – kód proteinu. Z toho vyplývá vztah obou databází, neboť kódování v rámci HGNC udává jednoznačný identifikátor pro daný gen (tj. HGNC:2597), a kódy z NCBI RefSeq tuto informaci rozšiřují. Představují totiž kódování v rámci více oblastí daného genu. Je také důležité zmínit dvě další významné databáze, které v rámci našeho popisu uvádíme níže. Jedná se o databázi vytvářenou HGVS a databázi NCBI dbSNP, které mají ke kódování NCBI RefSeq důležitý vztah, a které dále rozšiřují informace o daném genu.

2.4 Locus Reference Genomic sequences (LRG)

LRG [5] je paralelní databázi k databázi NCBI RefSeq. Systém LOINC (který je stručně popsán v kapitole 4) u příslušných položek uvádí, že mohou být kódovány buď hodnotami z databáze NCBI RefSeq nebo z databáze LRG. Web je dostupný z adresy *https://www.lrg-sequence.org*, databázi LRG lze nalézt ke stažení na webu *http://www.lrg-sequence.org/data/#lrg-data*. Data je možné stáhnout ve formátu XML nebo ve formát FASTA, který obsahuje přímo zápis genu, tj. ACGT sekvenci. Stažen je komprimovaný soubor (ZIP), který pro jednotlivé identifikátory obsahuje příslušející XML soubory. Další možností, kterou LRG nabízí, jsou tzv. Summary data, která jsou dostupná v textovém formátu TXT. Tato data obsahují kompletní přehled všech identifikátorů (kdežto v XML formátu je pro každý identifikátor určen jeden soubor). Tento přehled obsahuje také položku HGNC_SYMBOL, což je zkratka genu v systému HGNC, nikoli kód v rámci systému HGNC. XML soubory obsahují větší množství informací a jsou komplexněji uspořádány (hierarchie). Obsahují také podrobné mapování kódů LRG do databáze NCBI RefSeq. Příkladem kódu LRG je zápis LRG_1:g.8463G>C, nebo LRG_1p1:p.Gly191Ala.

2.5 NCBI Single Nucleotide Polymorphism Database (NCBI dbSNP)

NCBI dbSNP [6] je databáze jednonukleotidových polymorfismů, která pro ně zavádí kód s prefixem rs. Toto kódování tvoří v mnoha návazných systémech a databázích základní klíč pro vyhledávání a lze ho považovat vedle označení genu za nejdůležitější kódování v reprezentaci genetického vyšetření. Nicméně nikoliv za postačující. V mnoha případech bude nutné znát podrobnější informace představované databázemi NCBI RefSeq a HGVS. Databáze je dostupná na webu *https://www.ncbi.nlm.nih.gov/snp/*. Vzhledem k tomu, že NCBI zaštiťuje velké množství databází je dbSNP velmi dobrým integrátorem informací genetického záznamu. Pro rs kód obsahuje kromě genu i všechny příslušné referenční sekvence s HGVS kódy, ale i přehled studií v závislosti na populaci, odkazy na klinickou signifikanci (ClinVar, viz níže) nebo přehled publikací. Vzhledem k příslušnosti do NCBI portfolia je k dbSNP přístup přes API Entrez eUtils eSearch.

2.6 Human Genome Variation Society (HGVS)

HGVS [7] předkládá další možnou nomenklaturu pro popis variant sekvencí. Databáze je dostupná na webu *https://varnomen.hgvs.org/*. Zde je důležité zmínit, že HGVS má vedle sběru, dokumentace a bezplatné distribuce informací o genetických variancí přidružené i příslušné známe klinické informace. V rámci reprezentace genetické informace představuje rozšíření záznamu předchozích databází. V rámci systému NCBI je při rozkliknutí každého kódu typu rs# k dispozici výčet možných názvů HGVS – celková podoba kombinuje kódy z NCBI RefSeq, které jsou doplněny příslušným rozšířením HGVS, např. NC_000002.11:g.38298203C>G nebo NM_000104.3:c.1294C>G. Výhodou je možné odvození na základě propojení databáze HGVS a databáze SNP. Kódy HGVS systému mají svou vnitřní logickou strukturu a lze z nich určit informace o změnách jednotlivých sekvencí. Přehled pravidel pro tvorbu kódů lze nalézt na webu *http://www.hgvs.org/mutnomen/examplesDNA.html*. My si dovolíme tato pravidla zde uvést ve stručné formě. Je definováno sedm základních změn v sekvencích. Uveďme několik příkladů reprezentace změn v kódu substitution – např. g.423G>C, deletion – např. g.692_694delGAC, duplication – např. g.692_694dupGAC, insertion – např. g.451_452insGAGA, inversion – např. g.1077_1080inv.

Těchto sedm změn je vždy svázáno s oblastí genomové sekvence, které se týká. Pravidla pro specifické určení změny jsou dohledatelná v tabulce na *http://www.hgvs.org/mutnomen/examplesDNA.html#intro*. Na webu *http://www.hgvs.org/mutnomen/examplesDNA.html#genera* jsou také definována pravidla, která platí pro sestavení kódu, jakožto kombinace typu změny a její. Zde je třeba ještě dodat, že na změny lze pohlížet z úrovně genomové sekvence i z úrovně kódování DNA sekvence (případně i z hlediska proteinové sekvence).

3 Interpretace genetického vyšetření

Pokud mají být informace v rámci genetické karty dále sdíleny a následně využívány, je k těmto informacím nutné získat i interpretaci na klinické úrovni. V řadě případů tvoří tuto interpretaci expert na úrovni laboratoře a lékař již pracuje s touto interpretací. Existuje také ale několik databází určených ke sběru a distribuci známých propojení genetické a klinické informace. Podívejme se nyní na tři takové databáze, které jsou v rámci našeho záměru velmi důležité.

3.1 PharmGKB

Jako první uvádíme pro nás nejdůležitější z databází a tou je PharmGKB [8], která je k dispozici na webu *https://www.pharmgkb.org/*. Finančně je podporována NIH/NIGMS a provozuje ji Stanfordská univerzita. Databáze PharmGKB obsahuje a shromažďuje hlavně farmakogenomická data, která zahrnují klinické informace o potenciálně klinicky významných asociacích mezi geny a léky (a také chorobami) jakožto vztahy mezi genotypem a fenotypem podstatné pro dávkování léků. Tyto vztahu jsou v rámci databáze označeny jako klinické anotace. Shrnuje také informace o důležitých farmakogenomických genech, vztazích mezi genetickými variantami a léky a metabolickými procesy. Přehled interpretací je vázaný na gen a nejčastěji rs# kód z dbSNP (u některých genů je použita hvězdičková notace) ve spojení s lékem, chorobami a druhem popisované interakce jako je účinnost léku, dávkování nebo toxicita. Změny těchto interakcí jsou pak slovně popsány pro jednotlivé genotypy (tento popis představuje fenotyp). Uveďme příklad pro variantu rs4244285 genu CYP2C19 je definována klinická anotace pro účinnost amitriptylinu. V rámci anotace je konstatováno, že pro ty s genotypem AA nebo AG mají snížený metabolismus (zvýšenou koncentraci v plazmě) oproti genotypu GG. V tomto případě je v rámci interpretace i informace, že mohou hrát roli jiné genetické faktory ovlivňující dávkování a měly by tudíž být brány na zřetel.

Velmi důležitým parametrem přiřazeným každé klinické anotaci je level of evidence, tedy definovaný stupeň důvěryhodnosti dané anotace. Tento parametr je definován v rámci PharmGKB (*https://www.pharmgkb.org/page/clinAnnLevels*), kde evidence na úrovni 1A (schváleno odbornou společností, zařazeno do guideline) je nejlepší a evidence na úrovni 4 je nejhorší (případové studie, nesignifikantní nebo in vitro studie).

3.2 SNPedia

Databáze SNPedia [9] je přehledová databáze, zaměřená na

jednonukleotidové polymorfismy (SNPs) dostupná na *https://www.snpedia.com/index.php/SNPedia*. Propojuje polymorfismy reprezentované rs kódem s literaturou v databázi PubMed pro specifické pro nějaké riziko (např. cukrovky 2. typu, obezitou, atd). Toto propojení lze brát jako objektivní informaci a záleží na kvalitě samotných studií. SNPedia však přidává vlastní agregující parametr, který v některých případech subjektivně vystihuje míru zájmu o dané propojení genotypu a klinické informace. Uveďme příklad – 0 (nejnižší hodnota) je v rámci SNPedia vykládána, jako že o genomu není nic zajímavého známo – není propojení s klinickou informací. Magnitude 3 je definována tak, že informace takto označená je pravděpodobně dostatečně zajímavá. Magnitude 10 je definována jako signifikantní informace. V rámci databáze se nazývá Magnitude a jeho definice je dostupná na *https://www.snpedia.com/index.php/Magnitude*. Tento přístup není zcela rovnocenný přístupu PharmGKB pro náš systém, avšak je definován i pro vztahy genotypů a jiných klinických informací, než pouze lékových, jako je to u PharmGKB.

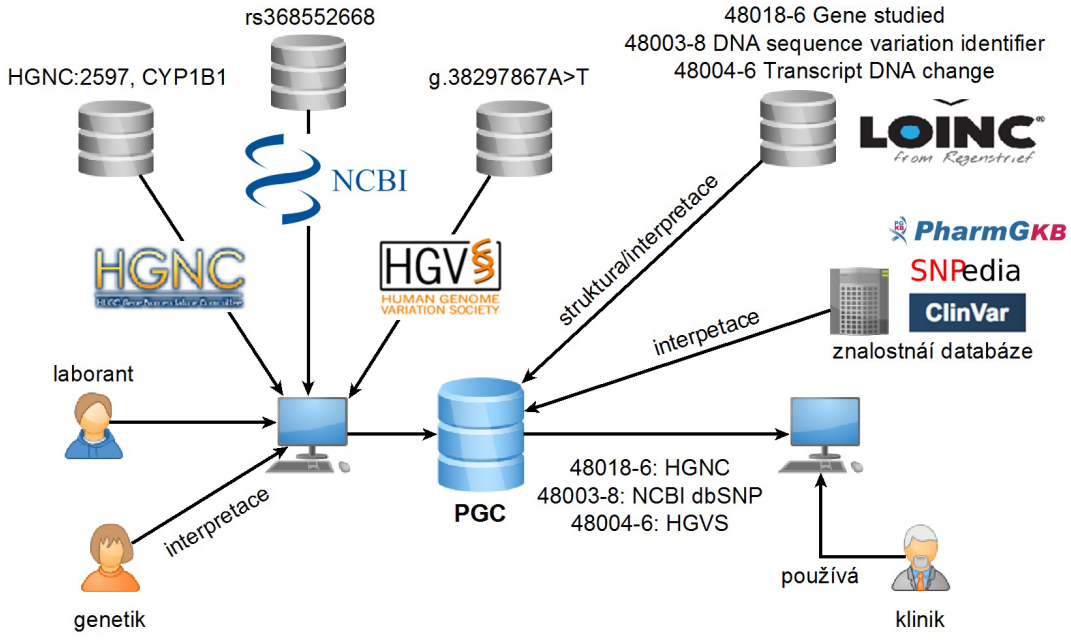
3.3 ClinVar

Poslední z popsaných interpretačních databází je odkazována databáze ClinVar (*https://www.ncbi.nlm.nih.gov/clinvar/*) [10] [11], která je provozována v rámci NCBI. Proto je přímo propojena s dbSNP, na které jsou přímo odkazy a klinické informace v ClinVar. ClinVar je stejně jako obě předchozí databáze svázán s klinickými informacemi přes genetických variací, ale na rozdíl od předchozích dvou databází je pro něj hlavním kódováním systém HGVS. Nicméně vzhledem k provázání s dbSNP a s ohledem na fakt, že tato spojuje kódování polymorfizmů pomocí rs kódů s doplněním kódů NCBI RefSeq a právě HGVS, není problém přes vyhledávač ClinVar vyhledávat i pomocí rs kódů. Avšak v samotné databázi ClinVar je např. pro rs kód rs12255372 je např. jeden ze dvou deklarovaných záznamů NM_001146274.2(TCF7L2):c.552+9017G>T. Tento výsledek genetické analýzy je svázán s rizikem diabetu 2. typu a stejně jako v předchozích případech je tento závěr učiněn a podložen studiemi, které jsou v rámci ClinVaru odkazovány (opět je to odkaz do databáze PubMed). Systém ClinVar umožňuje získání dat z ftp nebo v rámci Entrez eUtils API služeb NCBI.

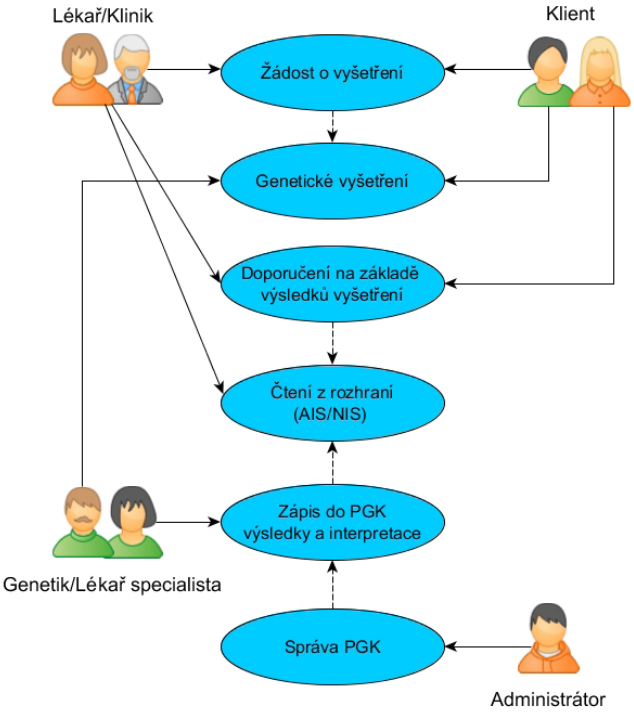
4 Personální genetická karta

Pro účely projektu jsme vytvořili návrh možného uspořádání systému, který budeme nazývat Personální genetickou kartou. Jelikož je genetické vyšetření bráno z hlediska reprezentace v nomenklatuře jako specifický druh laboratorního vyšetření, je i náš návrh více či méně reprezentací laboratorního vyšetření. Avšak s několika odlišnostmi. Přehledové schéma systému je zobrazeno na obrázku 2.

Jak je z obrázku 2 patrné, vstupy Personální genetické karty lze rozdělit do čtyř částí. První část tvoří komponenty LOINC, které tvoří terminologický základ pro obsah karty. Přesné uspořádání datového modelu není tolik důležité a pravděpodobně bude optimalizováno na co nejlepší výkon systému. Podstatou je řízená struktura, která umožňuje integritní ukládání údajů v jasném kontextu. Druhou částí je soubor databází, které definují obor hodnot v rámci výsledků genetického vyšetření, které jsou popsány výše v kapitole 2. Jejich použití je definováno i v rámci LOINC a jsou tedy v dané struktuře jasně ukotveny. Částečně může být obor hodnot doplněn i některými součástmi samotného LOINC a případně mohou být využity některé z konceptů systému SNOMED. Třetí důležitou částí je vstup ze strany znalostních databází, které obsahují interpretace genetických informací validní pro klinické prostředí. Tady rovnou zmiňme i část čtvrtou, kterou vstup z genetické laboratoře, která analýzu prováděla. V řadě zdrojů je jasně uvedeno, že interpretace genetického vyšetření by měla být provedena expertem, který umí posoudit validitu existujících zdrojů a výsledků a na jejichž základě rozhodne. Samozřejmě může v tomto ohledu využít existujících interpretačních databází, avšak ty by měly být následně v interpretaci citovány, aby bylo dohledatelné, na základě čeho byly závěry učiněny. Proces interpretace tedy není prostým přejímáním informací ze znalostních databází. Bez přítomnosti experta, který minimálně schválí systémem navrženou interpretaci získanou na základě předchozích případů, je zcela nezbytná. Klinik, který chce informace o genetickém profilu využít již pracuje s jejich dopadem na oblast, resp. pouze je informován o genetických důvodech, které ovlivnily navrhovanou medikaci. Postup procesu použití personální genetické karty je zobrazen v use-case diagram na obrázku 3.



Obrázek 2 – Schéma propojení informačních zdrojů pro vytvoření reprezentace a interpretace genetického vyšetření v rámci návrhu Personální genetické karty



Obrázek 3 – Use case diagram pro použití Personální genetické karty

5 Komunikace

V oblasti komunikace jsou pro oblast výměny genetických dat i s případnou interpretací nejlépe rozpracovány standardy v rámci HL7. Tyto jsou schopeny reprezentovat velmi rozsáhlou oblast informací a dat. V oblasti genetiky jsou primárně spojeny se systémem LOINC a tak jsou pro reprezentaci informací a výsledků obecně z laboratorních testů (a tedy i z genetických analýzy) velmi funkční. Navíc fakt, že tento přístup umožňuje definovat hodnoty v podstatě z libovolné databáze či číselníku, z něj činí velmi robustní nástroj pro sdílení a výměnu dat. Nicméně jeho rozsáhlost a komplexnost je zároveň jeho největší slabinou, neboť implementace funkčního mapování z informačního systému do zpráv není zdaleka triviální. V případě implementace je zde již zcela jistě nezbytný přístup mapování na úrovni objektů (v případě objektově orientované aplikace) a je nanejvýše vhodné využít existujících nástrojů a knihoven, reprezentujících standard ve zvoleném prostředí.

Zprávami, týkajícími se genomiky, se zabývá pracovní skupina HL7 Clinical Genomics Work Group. Hlavní dokumenty standardu v této skupině jsou popsány v [12].

5.1 HL7 verze 2

K dispozici jsou dva typy zpráv: zprávy z oblasti cytogenetiky [13] a zprávy z oblasti genetických variací [14][15]. Tyto dokumenty (specifikující obsah zpráv) jsou poměrně podobné, liší se v typech panelů LOINC, které používají. Zpráva typu Genetic Test Result Reporting [13][14] je definována hierarchicky (parent-child) propojenými panely nomenklatury LOINC. Tyto panely slouží jako vzory (template) pro příslušné zprávy. V obecném případě tyto definice panelů obsahují jeden kód, který identifikuje celý panel, a dále sadu kódů systému LOINC pro jednotlivé elementy potomků tohoto panelu. Potomkem může také být panel LOINC. Tyto panely se mohou opakovat a vytvářet tak struktury, které dokáží vyjádřit v reportu různé vzory. Každý potomek má definován datový typ, jednotky (měření), povinnost a seznam možných odpovědí (pokud je to přípustné). Důležitým pravidlem genetické zprávy je pravidlo, že všechny varianty by měly být popisovány na co možná nejzákladnější úrovni, tj. na úrovni DNA. Jednotlivé popisy by se vždy měly vztahovat k referenčním sekvencím, tj. referenčním sekvencím genomickým nebo kódové DNA.

5.2 HL7 Fast Healthcare Interoperability Resources (HL7 FHIR)

HL7 FHIR [16] je stále více populární standard v rámci rodiny HL7. Jeho obliba se dá přičítat několika faktorům. Za jeden z nejsilnějších považují autoři tohoto textu definování způsobu komunikace v rámci FHIR. To je určeno jako RESTfull API, které je v dnešní době standardním API v celé řadě aplikací a systémů. V rámci HL7 FHIR existuje třída Observation, která se zdá být vhodná pro použití u genetického vyšetření. Avšak pro určitá specifika definice pojmu Observation není doporučeno tuto třídu používat přímo. V rámci resources HL7 FHIR existují speciální rozšíření (extensions) již definovaných tříd. Z pohledu našeho záměru jsou nejzajímavější rozšíření Observation-genetics Profile [17] a DiagnosticReport genetics Profile [17], které definují hlavní údaje, které potřebujeme v rámci našeho záměru. Oba profily jsou opět navázány na terminologii LOINC. Oba profily obsahují dále několik komplexních tříd, jako jsou např. observation-geneticsVariant, observation-geneticsAllele nebo observation-geneticsInterpretation v Observation-genetics Profile. Tyto komplexní třídy umožňují přenášet informace v plném rozsahu daném v rámci použité terminologie.

5.3 HL7 CDA

Standard Clinical Document Architecture (CDA) je definován v rámci standardů HL7 verze 3, odvozený z Referenčního Informačního Modelu (RIM). Jestliže však v případě modelu RIM hovoříme o obecném modelu pro komunikaci ve všech oblastech (nejen lékařských) zdravotnictví, je CDA zaměřen výhradně na modelování pro účel výměny klinických informací. Dokumenty CDA mají jednoznačně definovanou strukturu (header a body) a implementační technologii (xml). Podrobnější informace k HL7 CDA R2 lze nalézt v dokumentu [18].

V oblasti genetických testů není použití CDA odlišné od ostatních oblastí laboratorních testů. U daného vyšetření je vytvořena struktura všech výsledků (observation), které by odpovídaly blokům OBX ve verzi 2. Stejně jako ve verzi 2 je hlavní terminologií pro prezentaci údajů komponenty ze systému LOINC. Avšak způsob vytváření zprávy je z podstaty verze 3 odlišný. Zejména proto, že struktura zprávy je zde komplexnější než ve verzi 2. Informace o způsobu vytváření zprávy jsou obsaženy v implementační příručce [19]. Základem jsou opět tzv. templates, tedy vzory či šablony, které jsou definovány pro jednotlivé části dokumentu. Jsou to hlavně Document Templates, Section Templates a Clinical Statement Tamplates. Tyto vzory jsou samozřejmě abstraktní a jejich instance je potřeba většinou doplnit nějakou výstupní hodnotou. To je reprezentováno strukturou ValueSet, kterou jsme definovali v našem návrhu genetické karty pacienta. Podrobný popis celé implementační příručky je dostupný v [19], a je velmi podstatné brát ho na zřetel a postupovat dle něj a samozřejmě splňovat všechny licenční podmínky.

6 Závěr

V tomto příspěvku jsme na základě úvahy a návrhu prostředků pro systém Personální genetické karty chtěli demonstrovat komplexnost využití různých slovníků, nomenklatur a terminologických slovníků při reprezentaci a interpretaci genetického vyšetření. Jak již bylo zmíněno, význam genetického vyšetření je velmi úzce svázán s účelem jeho použití, které je potřeba vždy zvažovat. Pokud bychom chtěli hledat přirovnání v rámci medicínské oblasti, asi nejbližší by z našeho úhlu pohledu byla

obrazová diagnostika, specificky rentgen. Tento příměr není myšlen ve smyslu využití nomenklatur a terminologií, protože to je značně odlišné. Je myšlen spíše v chápání významu a práce s takovým vyšetřením. Rentgenové vyšetření nechá lékař provést v okamžiku, kdy je to nutné, protože nechce pacienta zatěžovat dávkou RTG záření zbytečně. V případě genetického vyšetření není problém v účincích vyšetření, ale ve faktu, že takto získaná data budou již stále platná a musí být tedy využitelná, dobře chráněna a propojena s aktuální úrovní výzkumu v dané oblasti. Druhou úrovní tohoto přirovnání (pozn. každé přirovnání je vágní) je fakt, že lékař ne-radiolog bude využívat spíše popis snímku vytvořený radiologem. Tento přístup je stejný jako v případě že lékař-klinik využívá interpretace a závěry od genetika. Samozřejmě rozdíl je ve způsobu vzniku interpretací, kdy v genetické oblasti se bude jednat o závěry podložené empiricky. Tento způsob je však vhodný pro částečnou automatizaci celého procesu, kdy účelem je zjednodušit proces vytváření interpretací bez ztráty jejich validity. A právě oblast farmakogenetiky je pro tento záměr v této chvíli kruciální, neboť jednoznačně opodstatňuje genetické vyšetření a navazuje na něj v oblasti automatizované optimalizace administrace léků.

Poděkování

Práce byla podporována grantem Ministerstva průmyslu a obchodu ČR č. 2018FV30421 GENOMKIT – Progresivní technologie pro racionalizaci personalizované farmakogenomiky, nutrigenomiky a sportovní medicíny.

Literatura

[1.] Regenstrief Institute, Inc. Logical Observation Identifiers Names and Codes – LOINC®, [cit. 2020-02-17]. Dostupné z: www.loinc.org

[2.] Supporting interoperability of genetic data with LOINC, Jamalynne Deckard, Clement J McDonald, and Daniel J Vreeman1, Journal of the American Medical Informatics Assoc. 2015 May; 22(3): 621–627.

[3.] Yates B, Braschi B, Gray K, Seal R, Tweedie S, Bruford E. Genenames. org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017 Jan 4; 45(D1):D619-625.

[4.] Leary NA, Wright MW, Brister JR, Ciufo S, et. al. eference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-45

[5.] MacArthur JA et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants, Nucleic Acids Res. Volume 42, Issue D1, 1 January 2014, Pages D873–D878, doi: 10.1093/nar/gkt1198.

[6.] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.

[7.] Johan T. den Dunnen Raymond Dalgleish Donna R. Maglott, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update, HUMAN MUTATION, Vol. 37, No. 6, 564–569, 2016

[8.] M. Whirl-Carrillo, E.M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C.F. Thorn, R.B. Altman and T.E. Klein. „Pharmacogenomics Knowledge for Personalized Medicine" Clinical Pharmacology & Therapeutics (2012) 92(4): 414–417.

[9.] Michael Cariaso; Greg Lennon: SNPedia: a wiki supporting personal genome annotation, interpretation and analysis, Nucleic Acids Research 2011; doi: 10.1093/nar/gkr798

[10.] Melissa J Landrum,corresponding author Jennifer M Lee, Mark Benson, et al. ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 2018 Jan 4; 46(Database issue): D1062–D1067.

[11.] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, et al. ClinVar: public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 2016 Jan 4; 44(Database issue): D862–D868.

[12.] Standard Representation of Genomic Information, Yan Heras, Phd., Lantana Consulting Group, PPT presentation, 2013

[13.] HL7 version 2 Implemetation Guide: : Clinical Genomics; Fully Loinc--Qualified Cytogenetics Model, Release

[14.] HL7 version 2 Implemetation Guide: Clinical Genomics; Fully Loinc--Qualified Genetic Variatio Model, Release 1, ORU^R01, HL7 Version 2.5.1, April 2009

[15.] HL7 Version 2.5.1 Implemetation Guide: Orders and Observations: Interoperable Laboratory Result Reporting to EHR, Release 1, ORU-^R01, HL7 Version 2.5.1, November 2007 1, ORU^R01, HL7 Version 2.5.1, May 2011

[16.] HL7 International. HL7 Fast Healthcare Interoperability Resources Specification ‚FHIR™', Release 1. 2014-02-02. Archived from the original on 2014-12-28. [cit. 2020-02-17].

[17.] HL7 FHIR Genomics Implementation Guidance, cit. 2020-02-17]. Dostupné z: https://www.hl7.org/fhir/genomics.html

[18.] Robert H. Dolin, MD, Liora Alschuler, Sandy Boyer, BSP, Calvin Beebe, Fred M. Behlen, Phd, Paul V. Biron, Amnon Shabo (SHVO), Phd. HL7 Clinical Document Architecture, Release 2, 2006

[19.] Implementation Guide for CDA Release 2 Genetic Testing Report (GTR), Draft Standard For Trial Use, September 2012

Kontakt

Michal Huptych
ČVUT v Praze, CIIRC
Jugoslávských partyzánů 1580/3
160 00 Praha 6
tel: +420-22435-4168
e-mail: michal.huptych@cvut.cz
https://www.ciirc.cvut.cz/