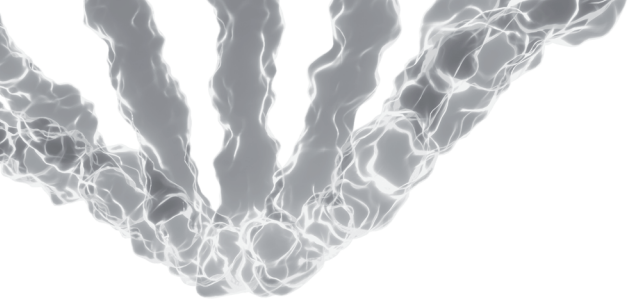
Towards solving the missing heritability in pharmacogenomics

Lee, M. van der



CHAPTER 8

General discussion and future perspectives

Introduction

Pharmacogenomics (PGx) is widely recognised as an important aspect in personalized medicine [1,2]. By analysing and interpreting one's genetic profile dose and drug adjustments can be made. In this way, one can strive to improve the safety and efficacy of drug treatments. Nonetheless, not all genetic variability in drug response can be explained with current PGx [3,4]. We hypothesized that this missing heritability can be explained by several factors. Firstly, rare and novel variants which are unaccounted for in routine PGx panels might play a role. Secondly, the complexity of pharmacogenes can result in an inability to unravel the genetic make-up of these genes. Thirdly, haplotype phasing is generally not taken into account in PGx. Fourthly, all genetic variants are currently summarized into one of four metabolic categories: poor metabolizers (PM), intermediate metabolizers (IM), normal metabolizers (NM) (previously EM) and ultra-rapid metabolizers (UM). However, enzyme activity is not a matter of 'on' or 'off', but is more of a continuous scale. Finally, the effect of a genetic variant on drug metabolism shows substrate specific effects. This substrate specificity can result in erroneous extrapolation of variant effects to the entire range of substrates. The development of novel technologies to determine one's genetic make-up is evolving rapidly, thereby providing opportunities for the field of PGx to address these issues [5-7]. This thesis explores these different factors and the role they play in the missing heritability.

Variant interpretation

Clinical PGx often focusses on single nucleotide variant (SNV)-panels or a selection of variants extracted from sequencing data. In chapter 3, we have investigated the feasibility of extracting such a PGx profile from existing diagnostic whole exome sequencing (WES) data [8]. For this study, a panel of clinically relevant variants was used. By extracting PGx variants from WES data, a clinically relevant profile similar to that obtained with panel testing can be generated. For 8 out of 11 genes, a clinically relevant PGx profile was generated, for *CYP2D6* copy number variants (CNVs) could not be determined and for *CYP3A5* and *CYP2C19* there was a lack of coverage on the intronic variants. While this approach does lead to a clinically relevant PGx profile, it also raises several concerns. Most importantly, regarding the selection of the best variants to include in a PGx panel. PGx variants can roughly be divided into three groups: common variants with high impact in well-known pharmacogenes, rare and common variants with a medium to low impact in genes associated with drug response and finally, rare (minor allele frequency (MAF) <1%) or novel variants in any genes associated with drug response (Figure 8.1).

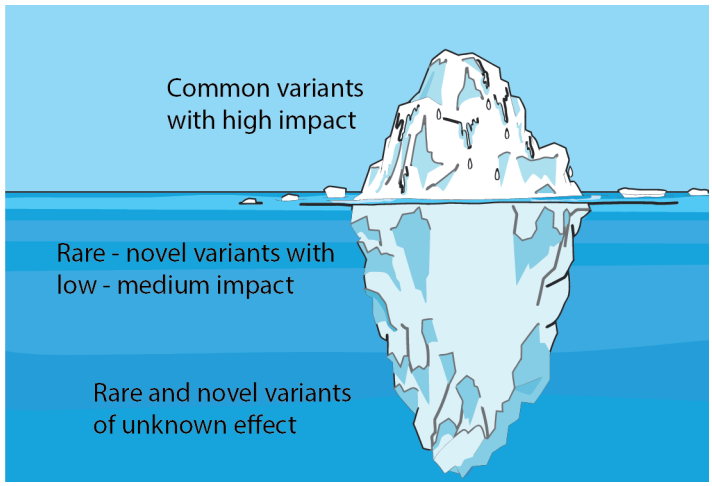


Figure 8.1: Types of pharmacogenomic variants
Three types of pharmacogenomic variants can be identified. 1) Common variants with high impact. These are often used in routine pharmacogenomic testing and are considered the low-hanging fruit in pharmacogenomics. 2) Rare and common variants of medium to low clinical impact, these variants have a small or unknown effect or they are too rare to be of clinical importance in the general population. 3) Rare and novel variants for which the effect is yet unknown, these variants make up the largest group.

The variants included in our repurposing study belonged to the first group, all were deemed clinically relevant and common. Unfortunately, due to the inherent limitations of WES data, we were unable to identify the selected intronic variants. For example, the variant rs12248560 (NC_000010.11:g.94761900C>T) which defines the *CYP2C19**17 allele is located upstream of *CYP2C19* and therefore not included in most exome sequencing panels. However, this variant occurs in approximately 20% of the European population and results in an increased metabolic capacity of CYP2C19, making it an important PGx variant. Given the high impact and frequency of this variant, it was decided not to report the CYP2C19 phenotypes using the *2 and *3 variants. Phenotype assignment without the *17 variant could give the illusion that all relevant variants were taken into account. The same approach was used for *CYP3A5* and *CYP2D6*, for which not all variants could be reported. While we have decided not to report the predicted phenotypes, other laboratories might determine CYP2C19 phenotypes without the *17 variant. Currently, there is no standardization of a minimally required variant panel, every laboratory can have their own selection of variants for which they test. This leads to discrepancies in phenotypes based on the test that is used or the laboratory that executed the genotyping [9]. The College of American Pathologists (CAP) has developed recommendations for a minimal required variant panel for *CYP2C9*, *CYP2C19* and *VKORC1* [10,11]. A recommendation for *CYP2D6*

is currently in the making. Before these recommendations are widely implemented as a standardized panel, care should be taken that it is clear which variants have been tested to allow for more detailed interpretation of PGx result and to assess the potential need for retesting including more variants.

With sequencing, variants throughout the entire genome (or exome) can be identified. This will result in many variants of the second group, variants with a low to medium clinical impact. These variants are not of high clinical value for the general population but can be meaningful for an individual patient. For example the variant characteristic for *CYP2D6*7* (*NC_000022.11:g.42127856T>G*) which has a frequency of <0.1% in the general population but does result in a non-functional allele for those who are carrier of this variant. To aid the use of these type of variants in clinical practice, tools have been developed to translate sequencing data into *-haplotypes and diplotypes [12-17]. If whole genome/gene sequencing data is used there is no longer a need for variant selection, as all locus have sufficient coverage and all variants can be taken into account. However, these tools have several limitations. Firstly, if the tools incorporate all known *-haplotypes, they will need to be updated continuously to keep up with the guidelines and research. Secondly, the different tools do not always provide the same results when the same data is analysed [12,13]. Twesigomwe et al. [13] showed that Stargazer [17], Astrolab [15] and Aldy [14] were not always in agreement in regards to the haplotypes called. Compared to the golden standard (75 GeT-RM samples), stargazer was accurate in 89% of the diplotype assignments, Aldy in 88% and Astrolab in 72% [13]. This leads to questions regarding their reliability and generalizability. While these tools can be a valuable addition to the field of PGx, it is important to understand how the tools work and how to interpret the results prior to applying them.

Variants belonging to the third group of variants, the rare and novel variants, are the most abundant in the general population. These rare variants are expected to be one of the major contributors to the missing heritability in PGx [18-20]. With the use of next generation sequencing, these variants can be detected. In our studies focussing on the clinical utility of sequencing (chapters 3 and 4) we did not include these rare variants as their clinical impact is currently often unclear. In chapter 5, where we focussed on the missing heritability of *CYP2D6*, we did detect rare variants as well. Out of the 216 variants identified, 155 (71.7%) had a MAF <1%. Of these 155, 35 were predicted to have an impact on enzyme function (slice site or missense variants) indicating the potential high impact of these rare variants. Unfortunately, the neural network approach used in this study is not able to accurately predict very low frequency variants (1 time occurrence). Due to their low frequency it is not feasible to obtain enough clinical samples to perform clinical studies focussing on the impact of rare variants. One solution to this problem is to study

clinical outliers [18]. In individuals with extreme phenotypes certain rare variants might be enriched. However, such an approach would probably identify mainly high impact rare variants and not the lower impact variants as their effect might be too subtle to cause extreme phenotypes. Another solution is the use of in silico prediction models. Well known in silico prediction tools such as SIFT, PolyPhen2 and MutPred are used in disease genetics to predict the impact of novel/rare variants [20,21]. However, disease genes have a higher sequence conservation due to the more strict evolutionary selection. These models are often based on a disease set and a healthy/neutral set [21]. Well known PGx variants are much more similar to neutral variants than they are to disease variants and are often part of the neural/healthy control set [21]. Therefore, as these tools cannot reliably be applied to pharmacogenes currently, the development of in silico variant effect predictors specifically targeted at PGx variants should be a priority.

Phasing and genetic complexity

Besides rare variants, structural variation, genetic complexity and haplotype phasing can also play an important role in PGx and contribute to the missing heritability. Long-read sequencing has been shown capable of resolving complex genetic regions as well as haplotype phasing without the need for computational approaches or pedigree information [6,22-28]. With haplotype phasing variants are assigned to their allele of origin leading to more accurate haplotype characterisations. If there are sufficient heterozygous variants in a locus, the different sequencing reads can be aligned based on the absence or presence of these heterozygous variants. The longer the reads, the higher the chance that there are sufficient variants to allow phasing. Approaches for haplotype phasing can be roughly categorized in three types. Firstly, direct phasing through sequencing, as described above. In chapter 4, the ability of long read sequencing to phase a panel of pharmacogenes was investigated. Out of the 100 genes, 73 were fully phased into haploblocks. For the 15 genes that were deemed fully complex, nine were resolved in haploblocks. Secondly, phasing can be performed using pedigree information where it can be inferred if a variant is inherited from the father or the mother. Finally, computation phasing based on linkage disequilibrium (LD). If two variants are in high LD they often occur together, based on this information it can be inferred what variants are located on the same allele. In clinical practice this LD approach is most often used as techniques for direct phasing are often too expensive and pedigree information is generally not known. This is also the approach that is used by the *-allele calling tools, Stargazer, Aldy and Astrolab. However, as we have shown in chapter 3, LD based phasing can lead to erroneous phenotypes. For *CYP2B6*, it is generally assumed that the *4 (NC_000019.10:g.41009358A>G) and the *9 (NC_000019.10:g.41006936G>T)

variant are in high LD and that, if they both occur, they are located on the same allele [29]. By using child-parent trio phasing we were able to show that in 1.5% of the cases the variants are not located on the same allele but on opposing alleles. While in this case there was no difference in predicted phenotype, these results highlight a potential problem in clinical PGx. With 42 variants (the panel used in chapter 3), there are still many individuals that carry only one variant in each pharmacogene. When this number increases, the odds of having multiple variants in the same locus is increasing as well and with that the chance of inaccurate phasing. As more variants are being discovered and used in PGx, the haplotype assignments become more complex and phasing more important.

In addition to haplotype phasing, long-read sequencing can also characterise variants in complex regions. The majority of pharmacogenes is at least partially complex (containing repetitive sections or showing high sequence homology) [30]. This complexity can lead to loss of function in the case of (partial) deletions and hybrid formations. Moreover, it can lead to an inability to detect SNVs if the region is too complex to characterize. While the costs for whole genome long-read sequencing are still very high, the costs for panel sequencing are substantially lower [31]. For short-read sequencing multiple panels for pharmacogenes have already been developed and shown valuable in clinical practice and in research [32,33]. For long-read sequencing such a panel is not yet available but would be a great opportunity for future developments. In chapter 4, the ability of long read sequencing to solve the complexities in such a panel was investigated. Besides haplotype phasing, long-read sequencing also resulted in accurate SNV calls even in complex genomic regions (recall and precision >98%). Moreover, the long-reads made it possible to discriminate between reads originating for highly similar genes, such as the *CYP2D6* gene and the *CYP2D7* pseudogene. When it is not clear if a specific sequencing read originates from the pharmacogene or from a neighbouring pseudogene, this might lead to incorrect phenotype assumptions.

Phenotype categorization

Ultimately, the individual variants, structural complexity and inferred haplotypes phasing are used to predict the enzymatic capacity of the pharmacogenes and thereby predict drug response. The CPIC and DPWG PGx guidelines use of a four category system for the majority of pharmacogenes. Namely, the poor (PM), intermediate (IM), normal (NM) and ultra-rapid metabolizers (UM). However, enzyme activity is not categorical (no function, decreased, normal or increased), but more a continuous scale. This effect can also be seen in the wide spread of activity within one metabolic category. With individuals genetically characterized as NMs, displaying enzymes activities similar to that of UMs or even PMs [34,35].

In chapter 7, we investigated the origin of variation in CYP2D6 phenotype within the scope of clinical PGx. In this study, two cohorts of patients treated with tamoxifen and one cohort of patients treated with venlafaxine were included [36-38]. As both tamoxifen and venlafaxine are primarily metabolised by CYP2D6, their rate of metabolism can serve as a proxy for CYP2D6 enzyme activity. Specifically we used the most CYP2D6 specific metabolic step, for tamoxifen this was the metabolic ratio (MR) of ln(endoxifen/desmethyltamoxifen) and for venlafaxine the MR ln(desmethylvenlafaxine/venlafaxine) was used. The variability within one category was highest in the groups with a predicted IM or NM phenotype. This is to be expected as these groups consist of a wide range of combinations of decreased and normal activity alleles. Most interestingly, it can be observed that, for tamoxifen metabolism, the combination of two decreased alleles (0.5+0.5) results in a lower metabolic capacity compared to the combination of an active and null-allele (1+0), despite the same overall gene activity score (1.0). One potential explanation for this is that not every decreased activity allele results in 50% residual activity. In fact, most alleles have been shown to have lower residual activity [39-41]. This in turn, results in a lower overall activity when two of these alleles are combined. Thoren et al. showed similar results with a higher activity in the individuals with one active allele compared to the combination of two decreased activity alleles. These results suggest that reclassification of allele combinations can help to decrease some of the variability. For example, assigning a IM status to the combination of 0.5+0.5 and NM status to the combination of 1+0 can already improve the explained variability. Nonetheless, a large amount of variability within one category is still unexplained, even when using the allele activity score combinations.

In chapter 5 we have investigated a method for phenotype predictions on a continuous scale. In this study, the same three cohorts as in chapter 7 were used. To fully characterise the complex *CYP2D6* locus, long-read, amplicon-based sequencing was used. To develop the model for haplotype independent phenotype prediction a neural network to predict the CYP2D6 enzyme activity (expressed as the MR) was trained based on 77 variants in the primary cohort of 561 individuals. This led to an increase in explained variability from 54% with the conventional categorical model to 79% with the neural network predictive model. For the second tamoxifen cohort (CYPTAM-BRUT, n=167) the explained variability increased from 35% to 66%. The CYPTAM-BRUT cohort contained data regarding CYP2D6 inhibitor use. In the results, it was observed that the inhibitor users had a higher predicted MR compared to the observed MR, indicating the impact of inhibitor use. For the venlafaxine cohort (n=69) the explained variability increased from 55% to 64%, which is a smaller increase potentially caused by substrate specific effects.

Simultaneous to the development of our model, McInnes et al. developed a similar model based on transfer learning [42]. This approach also showed a clear increase in

explained variability (R^2=0.712, for predicting the metabolic activity of *-alleles). With transfer learning, a high level network is first developed and trained based on high level data. Then, the contribution of each node in this high level is transferred to a new model which is further refined using detailed data. In the *CYP2D6* model from McInnes et al., the model is first trained on the *-haplotypes and their assigned activities. Then, clinical data is used to develop the refined prediction model within the existing *-haplotype model [42]. The benefit of transfer learning is that the sample size required is lower as the model is pretrained on known associations. The downside of this approach is that the assumption is made that the basis of the models is the same. In this case, that the *-haplotype activities are also accurate for the clinical data.

Both the model we presented and the model from McInnes et al. use a continuous scale to predict the CYP2D6 enzyme activity. This seems to result in a clear increase in explained variability. The success of these models clearly highlights the role that phenotype summarisation plays in the missing heritability. However, a neural network model can only learn patterns it observes and is not able to predict the impact of novel variants it has not yet seen. For novel variants this would mean that they have to occur at least once in the training dataset for the model to be able to predict their effect. Moreover, if two variants are in complete LD, meaning they do not occur separately, the model cannot identify the causal variant. This means that large sample sizes are needed to predict the effect of rare variants.

In addition to predicting the overall enzyme activity, our model was also capable of predicting individual allele activities on a continuous scale. These allele activity scores were combined in a neural network based combiner model to predict the final enzyme activity. Interestingly, when adding the two allele activity scores in the conventional way (allele 1 + allele 2) the explained variability was lower (73% for the main CYPTAM cohort) than when the combiner model was used (79%). By using the combined model, it was observed that when there was a combination of one inactive allele with one (partially) active allele the overall activity was higher than when there was a combination of two half function alleles, similar as was seen in chapter 6. However, this effect was seen for all combinations of allele activities (for example 0.6+0 >> 0.3+0.3). This suggests that there might be another mechanism at play which compensates an inactive allele but not a decreased activity allele.

Substrate specificity

The current PGx phenotype predictions are gene oriented and assume the same effect of a variant on any of the substrates. However, not every substrate is effected equally by a particular variant, as was also observed in chapter 5 with regard to venlafaxine. Aberrations of expected effect on drug metabolism, can at least partially be explained

by this substrate specificity. In chapter 6, we further investigate this effect in *CYP2D6* by comparing results from in vitro experiments. We found that there are many studies investigating in vitro activity of CYP2D6 substrates. However, only a small number of studies compare multiple substrates and variants. Moreover, large differences between in vitro studies significantly complicates the ability to quantify the effect of each variant-substrate combination [43-45]. Nonetheless, indications of substrate specific effects for several *-haplotype–drug combinations were identified, predominantly in *CYP2D6**17. Here it was observed that the *CYP2D6**17 activity was higher for debrisoquine than it was for dextromethorphan and bufuralol. Interestingly, *CYP2D6**17 causes a decrease in the substrate access channel and debrisoquine is much smaller than the other substrates. This might indicate that debrisoquine is not affected by the decrease in size of the access channel due to its smaller size.

While the effect of substrate specificity is widely known and increasingly better understood, it is not yet incorporated into the clinical guidelines. To enable this, a system to assess or predict the individual drug-variant interactions. As mentioned above, AI approaches can be valuable in variant effect prediction. A valuable future development can be the combination of a variant effect prediction model with drug characteristics such as molecule size. By combining these two aspects, substrate specific effects can potentially be predicted and incorporated in clinical practice. This does mean that the use of AI based prediction models first need to be accepted and standard in the field of PGx. Moreover, this application of AI will also require large sample sizes which are not always available in PGx.

Future perspectives

Rare variants, genetic complexity, haplotype phasing, phenotype summarisation and substrate specificity all play a role in the missing heritability in pharmacogenomics. Most clearly, the combination of long-read sequencing with complete haplotype phasing and a continuous phenotype prediction models seems to decrease the missing heritability substantially. The results presented in this thesis show what the role of these different factors play in the missing heritability. However, beyond *CYP2D6* there are many more pharmacogenes which suffer from missing heritability. Moreover, many additional (non)-genetic factors can influence drug response. More research is needed to account for these additional factors and to explore the implications of our findings for other pharmacogenes.

Due to its high complexity and involvement in the metabolism of many drugs, the *CYP2D6* gene is often the 'proof-of-principle' gene in pharmacogenomic studies. As was also the case in this thesis. However, there are many more genes that deserve attention. In particular the

CYP3A locus is of interest. This locus is characterized by many rare variants which are also predicted to play a significant role in enzyme activity [18]. The *CYP3A* locus consists of the *CYP3A4*, *CYP3A5* and *CYP3A7* genes. These genes share approximately 80% of their sequence [4]. This high sequence homology produces the same type of problems as what is seen for pseudogenes. Moreover, approximately 90% of the metabolism of the CYP3A4 substrates erythromycin and midazolam is expected to be genetic [46], while only around 10% of the activity of CYP3A4 can be explained by current clinical PGx [4].

Beyond shifting the focus to different genes, it is also important to address the impact of genetic variants influencing gene expression. This can either be intragenic, as is the case with *CYP2D6**10 [47-49], but also extragenic, for example the hepatocyte nuclear factor HNF4α [50-52]. One main limitation of studying the expression pharmacogenes is the fact that the Cytochrome P450 enzymes are predominantly expressed in the liver. To obtain data regarding expression liver biopsies are needed to analyse protein content and mRNA content [53]. Recently, Achour et al. proposed a method to analyse mRNA from liver exosomes isolated from blood samples as a measurement for expression [53]. These exosomes are shredded by the liver and contain mRNA of the genes expressed in the liver, included Cytochrome P450 genes. They compared the mRNA inferred expression with liver protein levels from the same individuals. They showed a good correlation between inferred and true expression for the CYP-enzymes (R^2=0.50–0.75). This method might be a useful tool to investigate CYP-expression without the need for invasive liver biopsies and should be explored further.

Besides pharmacokinetic interactions, genetic variants in pharmacodynamic targets and in transporters can also lead to significant differences in drug response. However, these interactions are generally less understood. One clear reason for this is that pharmacokinetic interactions can be measured objectively by assessing drug levels. Pharmacodynamic interactions often rely on reported phenotypes, such as a lack of efficacy or adverse reactions which are generally more difficult to characterise and therefore more difficult to associate with genetic variants. The same holds true for genetic variation in drug transporters. Drug transport depends on the drug concentrations, which can be influenced by genetic variants in genes associated with drug pharmacokinetics. Moreover, the drug transporters themselves can also be affected by genetic variation which influences their capability to transport drugs across membranes. To date there are only a few transporters included in the Dutch PGx guidelines, namely the *SLCO1B1* transporter related to statin toxicity and *ABCG2* related to allopurinol.

While genetic variants play an important role in the activity of pharmacogenomic proteins, there are many more factors which influence drug metabolism and enzyme activity

and cause a mismatch between the observed and predicted phenotypes. In the field of PGx, these non-genetic factors influencing enzyme activity are collectively known as phenoconversion. Two major types of phenoconversion are concomitant drug use (drug-drug-gene interactions (DDGI)) and comorbidity related phenoconversion [54,55].

DDGIs were briefly discussed in chapter 5 in regards to the outliers seen in the individuals using CYP2D6 inhibitors in the CYPTAM-BRUT cohort. The complexity of these interactions makes them often overlooked in studies. Nonetheless, they are expected to play a significant role in the variability of drug metabolism. Many drugs are inhibitors or inducers for common metabolic enzymes [56]. A combination of PGx variants, inhibitors and a substrate can result in complex pharmacokinetics which is not easily explained by adjusting for only one of the factors. Recently, physiology based pharmacokinetic (PBPK) models have been to model DDGIs [57,58]. One of these PBPK models showed a good prediction of kinetics of CYP2C19 mediated metabolism under the influence of inhibitors. The accuracy of the models did, however, differ between the predicted phenotypes of CYP2C19 [57].

For disease related phenoconversion many factors can play a role. For example; inflammation, cancer, pregnancy and liver disease [50,55]. One recent example of phenoconversion due to comorbidities is the alteration of CYP-enzyme activity due to COVID-19 infection. In a patient with COVID-19 it was observed that their metabolism of clozapine significantly changed compared to the metabolism before the COVID infection. This lead to toxic concentrations of clozapine [59]. The role of inflammation on pharmacokinetics is well known but not yet well understood. More in vitro and in vivo research towards this mechanism is needed to eventually be able to include inflammation as one of the markers to predict the activity of drug metabolic enzymes.

As the number of drug targets, drugs and genetic variants identified quickly increases, we are no longer able to process all this data into meaningful dosing guidelines without the use of AI. The main strength of machine learning is that the models are capable of detecting complex patterns in the data with which they are provided and trained on [60-62]. The implications of AI rise beyond the prediction of the metabolic capacity of just one enzyme based on its genetic make-up, as we did in our study. The field of AI has advanced far enough to make it possible to incorporate many different factors. Models to recognise patterns on medical scans or to predict drug risks are available in many different medical disciplines [63-68]. Nonetheless, to enable the broad implementation of AI in PGx one large limitation needs to be resolved. AI models are data driven, meaning they process large amounts of data into (clinically) meaningful outcomes. Large sets of high quality data, which is represents the real world situation well, is needed to train these AI models. However, the field of PGx is suffering from a lack of detailed phenotype data. There is simply not

enough data available to study all (potential) gene-drug interactions with the help of AI pattern recognition. For pharmacokinetic interactions, drug and metabolite levels would be the most suitable type of data as it is least affected by other factors. Hospitals are the optimal location to collect the data required for future PGx studies. Many hospitals have laboratories that can perform therapeutic drug monitoring (TDM) to collect drug and metabolite levels as well as (pharmaco)genetic departments which can aid in the collection of genetic material and data. To address the lack of data in PGx, one option would be the routine collection of the above mentioned materials and data in biobanks.

While pharmacokinetic outcomes are highly suitable to determine enzyme activity, the vast majority of data collected in PGx studies is related to therapy outcome and drug response. Most commonly in the form of reported efficacy and adverse drugs reactions (ADRs). However, these parameters are influenced by many more factors than just genetics. Moreover, they are often self-reported and therefore more subjective that TDM data. This makes it more difficult to generate (AI based) models predicting the drug response based on ADRs, as it is difficult to distinguish between a true ADRs caused by PGx variants and other complaints which are not gene or drug related. The amount of data that is required to find the true gene-drug relations based on this type of data is significantly larger than that need for TDM based associations. Nonetheless, the ultimate aim of PGx is to improve drug treatment outcomes, making the collection of data on drug response more valuable in clinical practice. Drug response outcomes in the form of ADRs are used by the PREPARE study from the ubiquitous pharmacogenomics consortium (U-PGx), one of the largest PGx implementation studies. The aim of this study is to assess the impact of the implementation of PGx on the occurrence of adverse drug responses. The recruitment of approximately 7,000 subjects has just been completed and results are expected by the end of 2021 [69]. Given the large cohort size, AI models based on reported phenotypes could be developed using this data. These models could provide meaningful insight into the link between genetic variants and clinical outcomes instead of just the association between genetics and drug metabolism as we have done in this thesis.

Ultimately, AI models are ready for implementation in a research setting and, in the near future, in clinical practice. The most important thing to keep in mind is that AI approaches are data driven and therefore, the usefulness and applicability of these models rely on the quality of the training data.

As the examples above highlight once more, PGx and the prediction of drug response is complex. To predict the drug response for an individual and reduce missing heritability, many factors need to be taken into account: full genetic make-up of the genes involved, substrate specificity, drug-drug-gene interactions, inflammation and any other markers influencing the drug levels (e.g. kidney function, liver function). Incorporating all these

factors in a model goes beyond human capability. With the rapid development of AI and sequencing technologies, models that integrate all these factors and provide a physician with the best treatment option for each specific patient can be developed. With that, true personalized medicine using all patient characteristics becomes an achievable goal in the near future.

References

1. Roden, D.M.; McLeod, H.L.; Relling, M.V.; Williams, M.S.; Mensah, G.A.; Peterson, J.F.; Van Driest, S.L. Pharmacogenomics. *Lancet (London, England)* **2019**, *394*, 521-532, doi:10.1016/s0140-6736(19)31276-0.
2. Relling, M.V.; Evans, W.E. Pharmacogenomics in the clinic. *Nature* **2015**, *526*, 343-350, doi:10.1038/nature15817.
3. Matthaei, J.; Brockmoller, J.; Tzvetkov, M.V.; Sehrt, D.; Sachse-Seeboth, C.; Hjelmborg, J.B.; Moller, S.; Halekoh, U.; Hofmann, U.; Schwab, M., et al. Heritability of metoprolol and torsemide pharmacokinetics. *Clinical pharmacology and therapeutics* **2015**, *98*, 611-621, doi:10.1002/cpt.258.
4. Klein, K.; Zanger, U.M. Pharmacogenomics of Cytochrome P450 3A4: Recent Progress Toward the "Missing Heritability" Problem. *Frontiers in genetics* **2013**, *4*, 12, doi:10.3389/fgene.2013.00012.
5. van der Lee, M.; Kriek, M.; Guchelaar, H.J.; Swen, J.J. Technologies for Pharmacogenomics: A Review. *Genes (Basel)* **2020**, *11*, doi:10.3390/genes11121456.
6. Wenger, A.M.; Peluso, P.; Rowell, W.J.; Chang, P.C.; Hall, R.J.; Concepcion, G.T.; Ebler, J.; Fungtammasan, A.; Kolesnikov, A.; Olson, N.D., et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **2019**, *37*, 1155-1162, doi:10.1038/s41587-019-0217-9.
7. Gordon, A.S.; Fulton, R.S.; Qin, X.; Mardis, E.R.; Nickerson, D.A.; Scherer, S. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenetics and genomics* **2016**, 10.1097/fpc.0000000000000202, doi:10.1097/fpc.0000000000000202.
8. van der Lee, M.; Allard, W.G.; Bollen, S.; Santen, G.W.E.; Ruivenkamp, C.A.L.; Hoffer, M.J.V.; Kriek, M.; Guchelaar, H.J.; Anvar, S.Y.; Swen, J.J. Repurposing of Diagnostic Whole Exome Sequencing Data of 1,583 Individuals for Clinical Pharmacogenetics. *Clinical pharmacology and therapeutics* **2020**, *107*, 617-627, doi:10.1002/cpt.1665.
9. Rasmussen-Torvik, L.J.; Almoguera, B.; Doheny, K.F.; Freimuth, R.R.; Gordon, A.S.; Hakonarson, H.; Hawkins, J.B.; Husami, A.; Ivacic, L.C.; Kullo, I.J., et al. Concordance between Research Sequencing and Clinical Pharmacogenetic Genotyping in the eMERGE-PGx Study. *The Journal of molecular diagnostics : JMD* **2017**, *19*, 561-566, doi:10.1016/j.jmoldx.2017.04.002.
10. Pratt, V.M.; Del Tredici, A.L.; Hachad, H.; Ji, Y.; Kalman, L.V.; Scott, S.A.; Weck, K.E. Recommendations for Clinical CYP2C19 Genotyping Allele Selection: A Report of the Association for Molecular Pathology. *The Journal of molecular diagnostics : JMD* **2018**, *20*, 269-276, doi:10.1016/j.jmoldx.2018.01.011.
11. Pratt, V.M.; Cavallari, L.H.; Del Tredici, A.L.; Hachad, H.; Ji, Y.; Kalman, L.V.; Ly, R.C.; Moyer, A.M.; Scott, S.A.; Whirl-Carrillo, M., et al. Recommendations for Clinical Warfarin Genotyping Allele Selection: A Report of the Association for Molecular Pathology and the College of American Pathologists. *The Journal of molecular diagnostics : JMD* **2020**, *22*, 847-859, doi:10.1016/j.jmoldx.2020.04.204.
12. Caspar, S.M.; Schneider, T.; Meienberg, J.; Matyas, G. Added Value of Clinical Sequencing: WGS-Based Profiling of Pharmacogenes. *International journal of molecular sciences* **2020**, *21*, doi:10.3390/ijms21072308.
13. Twesigomwe, D.; Wright, G.E.B.; Drögemöller, B.I.; da Rocha, J.; Lombard, Z.; Hazelhurst, S. A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping. *NPJ genomic medicine* **2020**, *5*, 30, doi:10.1038/s41525-020-0135-2.

14. Numanagić, I.; Malikić, S.; Ford, M.; Qin, X.; Toji, L.; Radovich, M.; Skaar, T.C.; Pratt, V.M.; Berger, B.; Scherer, S., et al. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nature communications* **2018**, *9*, 828, doi:10.1038/s41467-018-03273-1.
15. Twist, G.P.; Gaedigk, A.; Miller, N.A.; Farrow, E.G.; Willig, L.K.; Dinwiddie, D.L.; Petrikin, J.E.; Soden, S.E.; Herd, S.; Gibson, M., et al. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. *NPJ genomic medicine* **2016**, *1*, 15007, doi:10.1038/npjgenmed.2015.7.
16. Lee, S.B.; Wheeler, M.M.; Patterson, K.; McGee, S.; Dalton, R.; Woodahl, E.L.; Gaedigk, A.; Thummel, K.E.; Nickerson, D.A. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genetics in medicine : official journal of the American College of Medical Genetics* **2019**, *21*, 361-372, doi:10.1038/s41436-018-0054-0.
17. Lee, S.B.; Wheeler, M.M.; Thummel, K.E.; Nickerson, D.A. Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences. *Clinical pharmacology and therapeutics* **2019**, 10.1002/cpt.1552, doi:10.1002/cpt.1552.
18. Ingelman-Sundberg, M.; Mkrtchian, S.; Zhou, Y.; Lauschke, V.M. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Human genomics* **2018**, *12*, 26, doi:10.1186/s40246-018-0157-3.
19. Kozyra, M.; Ingelman-Sundberg, M.; Lauschke, V.M. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. *Genetics in medicine : official journal of the American College of Medical Genetics* **2017**, *19*, 20-29, doi:10.1038/gim.2016.33.
20. Drogemoller, B.I.; Wright, G.E.; Warnich, L. Considerations for rare variants in drug metabolism genes and the clinical implications. *Expert opinion on drug metabolism & toxicology* **2014**, *10*, 873-884, doi:10.1517/17425255.2014.903239.
21. Li, B.; Seligman, C.; Thusberg, J.; Miller, J.L.; Auer, J.; Whirl-Carrillo, M.; Capriotti, E.; Klein, T.E.; Mooney, S.D. In silico comparative characterization of pharmacogenomic missense variants. *BMC genomics* **2014**, *15 Suppl 4*, S4, doi:10.1186/1471-2164-15-s4-s4.
22. Ameur, A.; Kloosterman, W.P.; Hestand, M.S. Single-molecule sequencing: Towards clinical applications. *Trends in biotechnology* **2018**.
23. Ardui, S.; Race, V.; Zablotskaya, A.; Hestand, M.S.; Van Esch, H.; Devriendt, K.; Matthijs, G.; Vermeesch, J.R. Detecting AGG Interruptions in Male and Female FMR1 Premutation Carriers by Single-Molecule Sequencing. *Human mutation* **2017**, *38*, 324-331, doi:10.1002/humu.23150.
24. Borràs, D.M.; Vossen, R.; Liem, M.; Buermans, H.P.J.; Dauwerse, H.; van Heusden, D.; Gansevoort, R.T.; den Dunnen, J.T.; Janssen, B.; Peters, D.J.M., et al. Detecting PKD1 variants in polycystic kidney disease patients by single-molecule long-read sequencing. *Human mutation* **2017**, *38*, 870-879, doi:10.1002/humu.23223.
25. Schüle, B.; McFarland, K.N.; Lee, K.; Tsai, Y.C.; Nguyen, K.D.; Sun, C.; Liu, M.; Byrne, C.; Gopi, R.; Huang, N., et al. Parkinson's disease associated with pure ATXN10 repeat expansion. *NPJ Parkinson's disease* **2017**, *3*, 27, doi:10.1038/s41531-017-0029-x.
26. Browning, S.R.; Browning, B.L. Haplotype phasing: existing methods and new developments. *Nature reviews. Genetics* **2011**, *12*, 703-714, doi:10.1038/nrg3054.
27. Laver, T.W.; Caswell, R.C.; Moore, K.A.; Poschmann, J.; Johnson, M.B.; Owens, M.M.; Ellard, S.; Paszkiewicz, K.H.; Weedon, M.N. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Scientific reports* **2016**, *6*, 21746, doi:10.1038/srep21746.
28. Snyder, M.W.; Adey, A.; Kitzman, J.O.; Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nature reviews. Genetics* **2015**, *16*, 344-358, doi:10.1038/nrg3903.

29. Bloom, A.J.; Martinez, M.; Chen, L.S.; Bierut, L.J.; Murphy, S.E.; Goate, A. CYP2B6 non-coding variation associated with smoking cessation is also associated with differences in allelic expression, splicing, and nicotine metabolism independent of common amino-acid changes. *PloS one* **2013**, *8*, e79700, doi:10.1371/journal.pone.0079700.
30. Lauschke, V.M.; Milani, L.; Ingelman-Sundberg, M. Pharmacogenomic Biomarkers for Improved Drug Therapy-Recent Progress and Future Developments. *The AAPS journal* **2017**, *20*, 4, doi:10.1208/s12248-017-0161-x.
31. Vilella, A. Next-generation-sequencing, v1.5.7. Availabe online: https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/htmlview?hl=en_GB
32. Bush, W.S.; Crosslin, D.R.; Owusu-Obeng, A.; Wallace, J.; Almoguera, B.; Basford, M.A.; Bielinski, S.J.; Carrell, D.S.; Connolly, J.J.; Crawford, D., et al. Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. *Clinical pharmacology and therapeutics* **2016**, *100*, 160-169, doi:10.1002/cpt.350.
33. Rasmussen-Torvik, L.J.; Stallings, S.C.; Gordon, A.S.; Almoguera, B.; Basford, M.A.; Bielinski, S.J.; Brautbar, A.; Brilliant, M.H.; Carrell, D.S.; Connolly, J.J., et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clinical pharmacology and therapeutics* **2014**, *96*, 482-489, doi:10.1038/clpt.2014.137.
34. van der Lee, M.; Allard, W.G.; Vossen, R.H.A.M.; Baak-Pablo, R.F.; Menafra, R.; Deiman, B.A.L.M.; Deenen, M.J.; Neven, P.; Johansson, I.; Gastaldello, S., et al. A unifying model to predict variable drug response for personalised medicine. **2020**, 10.1101/2020.03.02.967554 %J bioRxiv, 2020.2003.2002.967554, doi:10.1101/2020.03.02.967554 %J bioRxiv.
35. Jukic, M.M.; Smith, R.L.; Haslemo, T.; Molden, E.; Ingelman-Sundberg, M. Effect of CYP2D6 genotype on exposure and efficacy of risperidone and aripiprazole: a retrospective, cohort study. *Lancet Psychiatry* **2019**, *6*, 418-426, doi:10.1016/s2215-0366(19)30088-4.
36. Dezentje, V.O.; Opdam, F.L.; Gelderblom, H.; Hartigh den, J.; Van der Straaten, T.; Vree, R.; Maartense, E.; Smorenburg, C.H.; Putter, H.; Dieudonne, A.S., et al. CYP2D6 genotype- and endoxifen-guided tamoxifen dose escalation increases endoxifen serum concentrations without increasing side effects. *Breast cancer research and treatment* **2015**, *153*, 583-590, doi:10.1007/s10549-015-3562-5.
37. Sanchez-Spitman, A.; Dezentje, V.; Swen, J.; Moes, D.; Bohringer, S.; Batman, E.; van Druten, E.; Smorenburg, C.; van Bochove, A.; Zeillemaker, A., et al. Tamoxifen Pharmacogenetics and Metabolism: Results From the Prospective CYPTAM Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **2019**, 10.1200/jco.18.00307, Jco1800307, doi:10.1200/jco.18.00307.
38. Neven, P.; Jongen, L.; Lintermans, A.; Van Asten, K.; Blomme, C.; Lambrechts, D.; Poppe, A.; Wildiers, H.; Dieudonne, A.S.; Brouckaert, O., et al. Tamoxifen Metabolism and Efficacy in Breast Cancer: A Prospective Multicenter Trial. *Clinical cancer research : an official journal of the American Association for Cancer Research* **2018**, *24*, 2312-2318, doi:10.1158/1078-0432.Ccr-17-3028.
39. Ramamoorthy, Y.; Tyndale, R.F.; Sellers, E.M. Cytochrome P450 2D6.1 and cytochrome P450 2D6.10 differ in catalytic activity for multiple substrates. *Pharmacogenetics* **2001**, *11*, 477-487, doi:10.1097/00008571-200108000-00003.
40. Shen, H.; He, M.M.; Liu, H.; Wrighton, S.A.; Wang, L.; Guo, B.; Li, C. Comparative metabolic capabilities and inhibitory profiles of CYP2D6.1, CYP2D6.10, and CYP2D6.17. *Drug metabolism and disposition: the biological fate of chemicals* **2007**, *35*, 1292-1300, doi:10.1124/dmd.107.015354.

41. Hicks, J.K.; Swen, J.J.; Gaedigk, A. Challenges in CYP2D6 phenotype assignment from genotype data: a critical assessment and call for standardization. *Current drug metabolism* **2014**, *15*, 218-232.
42. McInnes, G.; Dalton, R.; Sangkuhl, K.; Whirl-Carrillo, M.; Lee, S.-b.; Tsao, P.S.; Gaedigk, A.; Altman, R.B.; Woodahl, E.L. Transfer learning enables prediction of CYP2D6 haplotype function. *bioRxiv* **2020**, 10.1101/684357, 684357, doi:10.1101/684357.
43. Hiratsuka, M. Genetic Polymorphisms and in Vitro Functional Characterization of CYP2C8, CYP2C9, and CYP2C19 Allelic Variants. *Biol Pharm Bull* **2016**, *39*, 1748-1759, doi:10.1248/bpb.b16-00605.
44. Hiratsuka, M. In vitro assessment of the allelic variants of cytochrome P450. *Drug metabolism and pharmacokinetics* **2012**, *27*, 68-84, doi:10.2133/dmpk.dmpk-11-rv-090.
45. Zanger, U.M.; Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther* **2013**, *138*, 103-141, doi:10.1016/j.pharmthera.2012.12.007.
46. Ozdemir, V.; Kalow, W.; Tang, B.K.; Paterson, A.D.; Walker, S.E.; Endrenyi, L.; Kashuba, A.D. Evaluation of the genetic component of variability in CYP3A4 activity: a repeated drug administration method. *Pharmacogenetics* **2000**, *10*, 373-388, doi:10.1097/00008571-200007000-00001.
47. Hanioka, N.; Okumura, Y.; Saito, Y.; Hichiya, H.; Soyama, A.; Saito, K.; Ueno, K.; Sawada, J.; Narimatsu, S. Catalytic roles of CYP2D6.10 and CYP2D6.36 enzymes in mexiletine metabolism: in vitro functional analysis of recombinant proteins expressed in Saccharomyces cerevisiae. *Biochemical pharmacology* **2006**, *71*, 1386-1395, doi:10.1016/j.bcp.2006.01.019.
48. Fukuda, T.; Nishida, Y.; Imaoka, S.; Hiroi, T.; Naohara, M.; Funae, Y.; Azuma, J. The decreased in vivo clearance of CYP2D6 substrates by CYP2D6*10 might be caused not only by the low-expression but also by low affinity of CYP2D6. *Arch Biochem Biophys* **2000**, *380*, 303-308, doi:10.1006/abbi.2000.1936.
49. Sakuyama, K.; Sasaki, T.; Ujiie, S.; Obata, K.; Mizugaki, M.; Ishikawa, M.; Hiratsuka, M. Functional characterization of 17 CYP2D6 allelic variants (CYP2D6.2, 10, 14A-B, 18, 27, 36, 39, 47-51, 53-55, and 57). *Drug metabolism and disposition: the biological fate of chemicals* **2008**, *36*, 2460-2467, doi:10.1124/dmd.108.023242.
50. He, Z.X.; Chen, X.W.; Zhou, Z.W.; Zhou, S.F. Impact of physiological, pathological and environmental factors on the expression and activity of human cytochrome P450 2D6 and implications in precision medicine. *Drug metabolism reviews* **2015**, *47*, 470-519, doi:10.3109/03602532.2015.1101131.
51. Ning, M.; Duarte, J.D.; Stevison, F.; Isoherranen, N.; Rubin, L.H.; Jeong, H. Determinants of Cytochrome P450 2D6 mRNA Levels in Healthy Human Liver Tissue. *Clinical and translational science* **2019**, *12*, 416-423, doi:10.1111/cts.12632.
52. Pan, X.; Ning, M.; Jeong, H. Transcriptional Regulation of CYP2D6 Expression. *Drug metabolism and disposition: the biological fate of chemicals* **2017**, *45*, 42-48, doi:10.1124/dmd.116.072249.
53. Achour, B.; Al-Majdoub, Z.M.; Grybos-Gajniak, A.; Lea, K.; Kilford, P.; Zhang, M.; Knight, D.; Barber, J.; Schageman, J.; Rostami-Hodjegan, A. Liquid Biopsy Enables Quantification of the Abundance and Interindividual Variability of Hepatic Enzymes and Transporters. *Clinical pharmacology and therapeutics* **2020**, *109*, 222-232, doi:10.1002/cpt.2102.
54. de Jong, L.M.; Jiskoot, W.; Swen, J.J.; Manson, M.L. Distinct Effects of Inflammation on Cytochrome P450 Regulation and Drug Metabolism: Lessons from Experimental Models and a Potential Role for Pharmacogenetics. *Genes (Basel)* **2020**, *11*, doi:10.3390/genes11121509.
55. Klomp, S.D.; Manson, M.L.; Guchelaar, H.J.; Swen, J.J. Phenoconversion of Cytochrome P450 Metabolism: A Systematic Review. *J Clin Med* **2020**, *9*, doi:10.3390/jcm9092890.

56. Indiana University, department of medicine The Flockhart Table. Availabe online: https://drug-interactions.medicine.iu.edu/Main-Table.aspx
57. Kanacher, T.; Lindauer, A.; Mezzalana, E.; Michon, I.; Veau, C.; Mantilla, J.D.G.; Nock, V.; Fleury, A. A Physiologically-Based Pharmacokinetic (PBPK) Model Network for the Prediction of CYP1A2 and CYP2C19 Drug-Drug-Gene Interactions with Fluvoxamine, Omeprazole, S-mephenytoin, Moclobemide, Tizanidine, Mexiletine, Ethinylestradiol, and Caffeine. *Pharmaceutics* **2020**, *12*, doi:10.3390/pharmaceutics12121191.
58. Wojtyniak, J.G.; Selzer, D.; Schwab, M.; Lehr, T. Physiologically Based Precision Dosing Approach for Drug-Drug-Gene Interactions: A Simvastatin Network Analysis. *Clinical pharmacology and therapeutics* **2021**, *109*, 201-211, doi:10.1002/cpt.2111.
59. Tio, N.; Schulte, P.F.J.; Martens, H.J.M. Clozapine Intoxication in COVID-19. *Am J Psychiatry* **2021**, *178*, 123-127, doi:10.1176/appi.ajp.2020.20071039.
60. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *The New England journal of medicine* **2019**, *380*, 1347-1358, doi:10.1056/NEJMra1814259.
61. Zou, J.; Huss, M.; Abid, A.; Mohammadi, P.; Torkamani, A.; Telenti, A. A primer on deep learning in genomics. *Nature genetics* **2019**, *51*, 12-18, doi:10.1038/s41588-018-0295-5.
62. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **2019**, *25*, 44-56, doi:10.1038/s41591-018-0300-7.
63. Schwalbe, N.; Wahl, B. Artificial intelligence and the future of global health. *Lancet (London, England)* **2020**, *395*, 1579-1586, doi:10.1016/s0140-6736(20)30226-9.
64. Jha, S.; Topol, E.J. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *Jama* **2016**, *316*, 2353-2354, doi:10.1001/jama.2016.17438.
65. Chang, H.Y.; Jung, C.K.; Woo, J.I.; Lee, S.; Cho, J.; Kim, S.W.; Kwak, T.Y. Artificial Intelligence in Pathology. *J Pathol Transl Med* **2019**, *53*, 1-12, doi:10.4132/jptm.2018.12.16.
66. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H. Artificial intelligence in radiology. *Nature reviews. Cancer* **2018**, *18*, 500-510, doi:10.1038/s41568-018-0016-5.
67. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nature reviews. Genetics* **2015**, *16*, 321-332, doi:10.1038/nrg3920.
68. Dias, R.; Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome medicine* **2019**, *11*, 70, doi:10.1186/s13073-019-0689-8.
69. van der Wouden, C.H.; Cambon-Thomsen, A.; Cecchin, E.; Cheung, K.C.; Davila-Fajardo, C.L.; Deneer, V.H.; Dolzan, V.; Ingelman-Sundberg, M.; Jonsson, S.; Karlsson, M.O., et al. Implementing Pharmacogenomics in Europe: Design and Implementation Strategy of the Ubiquitous Pharmacogenomics Consortium. *Clinical pharmacology and therapeutics* **2017**, *101*, 341-358, doi:10.1002/cpt.602.