Genetics and clinical characteristics to predict rheumatoid arthritis: where are we now and what are the future prospects?

Annette HM van der Helm-van Mil[†] *& Tom WJ Huizinga*

[†]*Author for correspondence*
Leiden University Medical Center, Department of Rheumatology, P.O. Box 9600, 2300 RC Leiden, The Netherlands
Tel.: +31 715 263 598;
Fax: +31 715 266 752;
AvdHelm@lumc.nl

In medicine, predicting disease development involves three major factors: the variability of the host, the characteristics of the disease-causing agent and the interactions among these factors (i.e., the disease process itself). This review focuses on the prediction of rheumatoid arthritis (RA) development in patients with undifferentiated arthritis (UA). Data from inception cohorts have revealed that, in approximately a third of patients presenting to a rheumatologist with recent onset arthritis, no diagnosis can be made, resulting in so-called UA. Although RA develops in a proportion of these patients, a substantial proportion spontaneously remits. The reason(s) for RA development as opposed to remission are as yet unknown. However, current epidemiological data on RA incidence rates are similar, with host (genetic) factors being an independent predictor of RA susceptibility. Also, human leukocyte antigen and protein tyrosine phosphatase nonreceptor 22 alleles have now been identified as risk factors in a number of populations, although most genetic factors involved are yet to be identified. Disease characteristics, such as the presence of anticyclic citrullinated peptide antibodies and erosions on x-rays, are also identified as being of high predictive value. The use of models that take into account both genetic and clinical characteristics to evaluate patient groups is important. In the future, the accuracy of these models in predicting, with 80% probability, the chance of progression from UA to RA should be established. Such prediction models will aid in determining the most suitable treatment for these patients.

A number of different factors are used in clinical prediction models. These include variability of the host, of the causative agent, in the disease process and, finally, the dynamics of the interaction that also takes time into account. Examples of the relevance of measuring pure host characteristics include the presence of the breast/ovarian cancer gene (*BRCA)-1/2* in families with a history of breast/ovarian cancer. The causative microorganisms in infectious diseases, such as in community-acquired pneumonia are examples of the relevance of the characteristics of the causative agent *in sensu strictu*. An example from rheumatology, when both host and causative agent determine susceptibility, is that certain microorganisms (e.g., *Chlamydia*) cause disease (reactive arthritis), particularly in hosts that are human leukocyte antigen (HLA)-B27 positive. Detecting differences in prognosis by studying the mode of interaction of the host and disease-inducing process include the study of diseased tissue characteristics, such as microarray studies in breast cancer. An example where the evolution of the disease over time is included in the determination of a response parameter is cervical abnormalities detected by a slightly abnormal cervical smear test. The value of this is that it is advised that the cervical smear test is repeated within 3 months to see whether natural regression of the abnormalities or progression to abnormalities indicative of pre-cancerous characteristics has occurred (Figure 1).

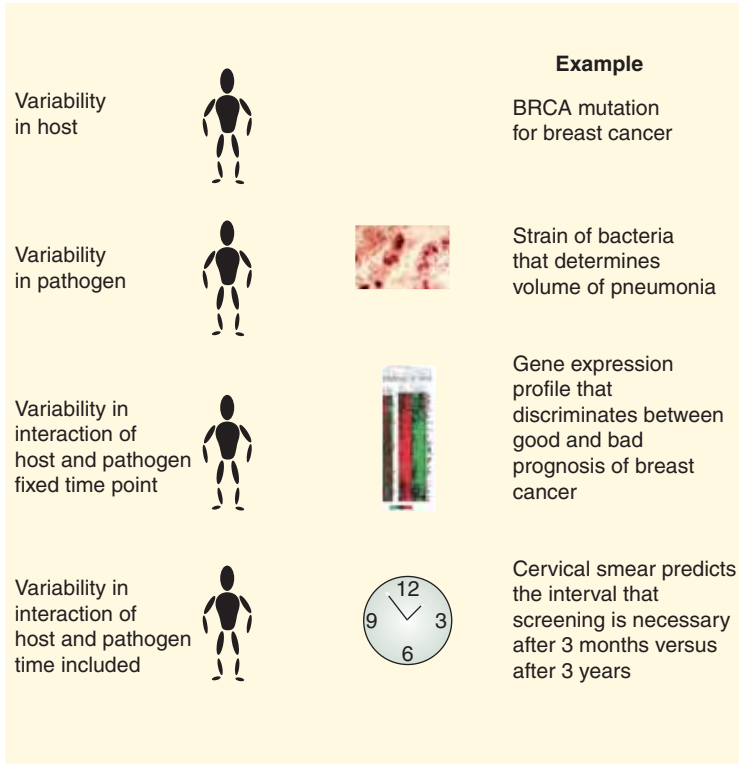
The outcome of a prediction model can be the development of disease or disease severity. In rheumatology, and particularly in rheumatoid arthritis (RA), physicians want to avoid morbidity and disability. Existing prediction models are therefore built to predict chronicity and erosiveness. Prediction models are of importance as they might help in treatment decisions. They may guide the choice of treatment options among wait and see, start with a relatively mild treatment or initiate aggressive treatment directly.

The current evidence for (early) treatment of RA is based on large trials with RA patients, in which RA is defined according to the American College of Rheumatology (ACR) criteria. Inception cohorts included patients in who, during the first visits with the current methodology, a diagnosis can be made directly (approximately 60% of patients). Approximately 40% of patients in inception cohorts have a form of arthritis in which no definite diagnosis can be made; these patients are

Keywords: anti-CCP, disease characteristics, genetics, HLA, incidence, prediction, PTPN22, rheumatoid arthritis, undifferentiated arthritis

future medicine

Figure 1. Different items in prediction models.



identified as having undifferentiated arthritis (UA). These UA patients can go into remission, develop RA or develop other conditions (Figure 2) [1]. At present, no data on the effects of treatment of UA patients are available. Early treatment of patients with UA that will develop RA might be beneficial, whereas treatment of the group that will remit spontaneously is potentially harmful. The spontaneous remission rate of UA patients is approximately 40% [1]. As current knowledge of the effects of RA treatment is based on patients with RA classification according to the ACR criteria, it will be helpful to have a model to predict RA development in UA patients. The prediction of RA is, in most cases, synonymous with the prediction of disease persistence as the remission rate of RA is approximately 10–18% [2,3].

This review focuses on the prediction of RA development. The following sections will review the evidence that pure host characteristics are informative, pure disease characteristics are relevant and also available evidence, highlighting that a combination of disease characteristics, host characteristics and natural course allows prediction. The applicability of prediction models is also determined by basic epidemiological rules and their predictive value is dependent on the prevalence of the disease in a given population. For the question regarding genetic testing, 'where we are now and what are the future prospects?', this is relevant because most current data on genetic factors describe the comparison between confirmed cases and healthy controls. These are relevant data to determine pathogenesis, but are difficult to interpret with regard to relevance in diagnostic testing in individual patients.

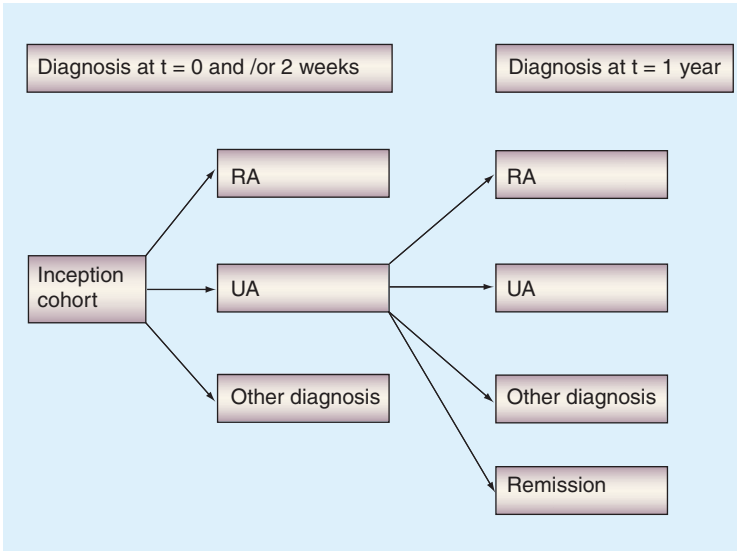
Variability in the host determines incidence of RA

This hypothesis has a number of relevant implications for RA. First, it assumes that the trigger or triggering events for RA are common and, therefore, that the opportunity to encounter these triggering events is not the limiting or determining factor. Second, it has implications as to whether RA is one disease or an assembly of truly different diseases. Third, it assumes that a number of DNA variants are associated with disease and that a DNA fingerprint might predict disease susceptibility.

Does RA have a common trigger?

The assumption that RA has a common trigger or a combination of commonly available triggers implies that these can be encountered in most populations. An argument in favor of this assumption is that RA is a worldwide condition, indicating that the triggers leading to RA must be available worldwide. RA incidence is age related, with a higher RA incidence with increased age. This age-related incidence of RA provides some evidence as to the number of triggers needed for RA development. Roberts-Thomson and colleagues studied population data obtained from the Australian Bureau of Statistics in order to assess the number of events necessary for RA development [4]. By computer modeling in which the age-specific incidence rates, the proportion of the population at risk and the age at onset are included, the number of random events that must occur for the disease to manifest (given a stochastic model) was calculated. This number varied somewhere between four and six events. In inception cohorts, the patients with UA at inclusion, who after 1 year of follow-up had persistent UA, were significantly younger than the UA patients that developed RA during the first year [1]. The difference in age between the groups that

Figure 2. Flow diagram of inception cohort.



RA: Rheumatoid arthiritis; UA: Undifferentiated arthritis.

do and do not progress to RA might reveal a difference in time period and subsequent chance to acquire sufficient numbers of triggering events.

RA has a variable incidence in different populations. In Pima Indians, the incidence rates in the same time periods were ten-times higher than in the Caucasian US population and five times higher than in the Japanese population [5]. This implies that the frequency of the events leading to RA is varied in different populations and/or that the host factors in different populations differ. The relative contribution of genetic or environmental factors is difficult to determine, but based on studies of populations that have migrated to different environments, it is likely that the majority of the differences in rates of RA in different populations can be explained by genetic factors [6]. Moreover, the differences in the frequency of the identified genetic risk factor for RA, that is, the HLA alleles encoding the shared epitope, associate with the frequency of RA in the respective populations [7]. The absence of geographical clustering of RA incident cases provides an additional argument that RA is caused by commonly available triggers [8]. Altogether, these data suggest that, with a modern lifestyle, a combination of triggers is common in a variety of cultures, but host characteristics determine whether these triggers can lead to RA.

Further understanding can be achieved by studying populations among which RA is not prevalent. A study in indigenous people (Aboriginals) in Australia found no paleopathological or ethnographical evidence to support the existence of RA before white settlement [9]. Similarly, in a rural Nigerian population, RA was not observed [10], indicating that either the common trigger was not present at that time or that these populations are protected genetically. Arguments for this last statement include the much lower frequency of HLA-D related (DR) alleles, which encode RA risk alleles in these populations. Careful studies have now identified RA in Aboriginals. However, in all Aboriginal RA patients, some evidence of prior inter-racial marriage was found. This indicates that genetic admixture is necessary for the development of RA. Yet, a contribution of changing lifestyles that is concomitant to racial admixtures cannot be excluded easily.

In a study from Minnesota that studied RA incidence rates from 1955 to 1995, the incidence rate fell progressively over the 4 decades of study, from 61.2/100,000 in 1955–1964 to 32.7/100,000 in 1985–1994 [11]. A Japanese study showed that the incidence rates also fell in Japan. Falling incidence rates over time have also occurred in other diverse populations, such as the Indians and Finnish [12]. There are several possible explanations for this decrease in RA incidence. Because it is apparent in various populations, the explanation is probably a factor that has an identical effect in all populations throughout the world in the birth cohorts from 1890 to 1950. It is proposed that this factor is a change in the population genome [13]. The explanation for this genetic drift is that, in previous times, human reproductive success was distributed unevenly, with a minority of fertile women giving birth to the majority of newborns. For example, in the 1912 Australian census, 50% of the children were the offspring of one in seven of the women [14]. However, in recent times, this predominance has decreased steadily since both fertile and less fertile women have contributed equally to the next generation. There are also other explanations for the decrease in RA incidence rates over time. Besides a real-time-dependent decline in RA, changing methodology in classification may also be important [15]. In addition to a decrease in RA incidence, a decrease in RA severity over time is also reported. This decline seems to be contributable to earlier and more aggressive treatment [16].

In summary, the overview of the studies of incidence rates of RA are compatible with the notion that host characteristics are the major factors that drive whether or not a patient will develop RA.

The authors suggest that most individuals nowadays will encounter those 4–6 triggering events and host factors are therefore the driving force to explain differences in incidence rates.

RA: one disease or an assembly of different diseases?

At present, RA is diagnosed formally when patients fulfill the criteria that were formulated by the ACR in 1987. Whether the patients that have RA, according to these criteria, all have the same disease – characterized by an identical pathogenesis – is questionable. Recently, it was observed in a European and American population that RA patients carrying antibodies to citrullinated proteins (anticyclic citrullinated peptide [CCP] antibodies) have an association with different genetic risk factors than patients lacking these antibodies. The shared epitope encoding HLA alleles only conferred risk to anti-CCP-positive and not anti-CCP-negative RA [17]. Anti-CCP antibodies are reported to have high disease specificity and are often present before the clinical presentation [18,19]; they are therefore thought to play a role in RA pathogenesis. The finding that the shared epitope alleles only correlate with anti-CCP-positive disease suggests strongly that RA patients with anti-CCP antibodies have differences in the pathophysiological pathway compared with RA patients that are anti-CCP-negative. This leads to the question: are anti-CCP-positive and -negative RA different disease entities with distinct clinical characteristics? In a recent study, RA patients with and without anti-CCP antibodies were compared extensively with regard to clinical characteristics. No differences were found in the characteristics on disease presentation between these two patient groups, including the age of disease onset, the type of initial symptoms, the distribution of initial symptoms, the presence and duration of morning stiffness and the number and distribution of painful or swollen joints [20]. From these data, it can be concluded that different pathophysiological pathways end in one phenotypical presentation of the disease. Specific characteristics of the host, such as the presence of anti-CCP antibodies, associate subsequently with the course of the disease.

Which potential genetic risk factors for RA are known?

The HLA class II molecules are the most powerful genetic factors recognized so far for RA, contributing to at least 30% of the total genetic effect. The HLA-DRB1 alleles **0101*, **0102*, **0401*, **0404*, **0405*, **0408*, **1001* and **1402* share a conserved amino acid sequence at positions 70–74 in the third hypervariable region of the DRβ1 chain. These residues constitute an α-helical domain forming one side of the antigen-presenting binding site. The 'shared epitope hypothesis' postulates that the shared epitope motif itself is involved directly in the pathogenesis of RA by allowing the presentation of an arthritogenic peptide. Extensive evidence exists showing associations between the shared epitope-encoding alleles and RA susceptibility. The presence of shared epitope-encoding alleles is associated with an odds ratio of approximately three to four to develop RA [21,22].

The second genetic risk factor is a risk allele of the hematopoietic-specific protein tyrosine phosphatase nonreceptor, (PTPN) 22. This allele was identified in 17% of North American Caucasian controls and 28% of RA patients, confering odds of approximately two to develop RA [23–26]. This allele changes the function of the protein that is a negative regulator of T-cell activation, leading to T cells with a lower threshold for T-cell activation. This mutation apparently leads to several autoimmune diseases since this mutation also confers risk for systemic lupus erythematosus (SLE), Type 1 diabetes and Graves disease [27,28].

Over recent years, an increasing number of single nucleotide polymorphisms (SNPs) associated with RA have been identified. Some results have not been replicated and some show different results in different populations. One genetic risk factor that is under investigation currently and seems to be associated with RA, diabetes and myocardial infarction, is major histocompatibility complex (MHC) class 2 transactivator (MHC2TA). This SNP associates with a lower expression of MHC molecules and, in a Swedish cohort of 1288 RA patients and 709 controls, this SNP conferred a 1.3 times higher risk of developing RA [29]. The findings on this SNP await replication. In Japanese patients and controls, an association between haplotypes (combinations of SNPs on one chromosome that tend to be inherited together) of the gene encoding peptidylarginine deiminase 4 (PADI4) and an increased susceptibility to RA was observed [30]. The RA-susceptible *PADI4* variant produces a more stable transcript than the nonsusceptible variant, implying increased production of PADI4 and, therefore, higher levels of citrullination by the RA-susceptible variant. Unless

increased citrullination occurred, the described PADI4 haplotypes did not correlate with (the level of) anti-CCP antibodies [30]. The association of PADI4 with RA is shown in the Japanese population. Data from Caucasians from France and the UK, however, showed no association between PADI4 haplotypes and RA [31,32].

Susceptibility genes can interact such that the resulting predisposition of carrying both genes is larger than the summed ratios of the individual genes. The presence of such interactions is important with regard to prediction. In 820 Japanese RA patients and 620 controls, risk for RA of 1.3 was identified for a risk allele in the organic cation transporter gene *SLC22A4* [33]. Intriguingly, the identified SNP affects the transcriptional efficiency of *SLC22A4 in vitro* by altering the binding affinity of a hematopoietic transcription factor, called RUNX1. A small but significant association was observed with the minor allele in the *RUNX1* gene. Importantly, homozygosity for both susceptibility alleles (*SLC22A4* and *RUNX1*) resulted in a high odds ratio of nine, indicative of a gene–gene interaction [33]. Recently, the effects of this *RUNX1* SNP were not found in a Caucasian population [34]. An SNP in the promoter region of *FCRL3* has been shown recently to be associated with RA susceptibility [35]. For a large number of other genes suggested to be relevant in the pathophysiology of RA, association was observed in only one study, without replication. These studies concerned β-adrenergic receptor gene SNPs, receptor activator of NF-κB ligand *(RANKL)*, anti-intercellular adhesion molecule *(ICAM)-1*, vascular endothelial growth factor *(VEGF)*, programmed cell death *(PDCD)-1* and interleukin *(IL)-1-RA* genes [36–40].

Besides genetic risk factors that confer a higher risk of developing RA, there are also genetic risk factors that protect from RA. This concerns particularly the HLA-DRβ1 alleles that encode for the amino acids DERAA (*DRB1*0103*, **0402*, **1102*, **1103*, **1301*, **1302* and **1304*). Interestingly, the HLA-DRB1 alleles can encode for different alleles with an opposite effect on disease susceptibility. The protective effect of the DERAA-encoding alleles is independent from the shared epitope-encoding alleles that have predisposing effects [22,41,42]. Both in the presence and absence of shared epitope-encoding alleles, the DERAA-encoding alleles confer significantly lower odds of 0.6 of developing RA [22].

In summary, the current knowledge of well validated genetic risk factors to be included in a DNA fingerprint is limited to *HLA-DRB1* and *PTPN22*. *HLA-DRB1* is estimated to account for 30% of the genetic component of this autoimmune disease [43], while the contribution of *PTPN22* is much smaller. Thus, a significant part of the genetic contribution is still to be identified. In a number of whole genome scans, many peaks of linkage have been identified [44]. In a study to estimate the number of true RA gene regions, which took into account both the heterogeneity of RA and the performance of a dense genome scan, 8 ± 4 regions (mean ± standard deviation [SD]) were found to be true positives and evidence for three additional regions was provided from covariate-based analysis [45]. One of those regions is the *HLA-DRB1* locus, meaning that at least 10 ± 4 additional genes will be identified each with a modest effect. Technical progress, such as SNP-based linkage analysis, has been demonstrated to allow loci to be defined more precisely [46]. The chance that this will lead to the identification of the majority of the genetic risk factors is larger if RA is caused by a dozen common genetic variants than if RA is the result of many rare mutations. Given the fact that *HLA* and *PTPN22* have already been identified, the authors speculate that RA is caused by a dozen common genetic variants. The statistical methods to evaluate many gene variants with disease status, as in candidate-gene case–control studies, are still in their infancy, especially for the low effect sizes of the individual disease loci and the occasionally low frequencies of the disease allele(s). The standard methods of evaluating the association of multiple markers with disease status are based on multimarker multivariate analyses. For such analyses, one typically uses logistic regression to test simultaneously the main effects (and possibly interactions) of multiple markers. For each marker, a covariate can be created, such as the number of rare alleles at each marker. When this type of coding is used in logistic regression, the resulting score statistic for each marker implies many degrees of freedom, implying that the overall model suffers from weak power. Moreover, complex models tend to overfit the data, stressing the necessity for replication in independent cohorts. Despite these difficulties, it appears that the genetic contribution to RA is approximately 50–60% [47]. This number is estimated

by variance component analysis in monozygotic and dizygotic twins [47]. This high percentage implies that measurement of genetic host characteristics is likely to have a role in a predictive test.

Which environmental risk factors for RA are known?

So far, smoking has been shown to be the only plausible environmental risk factor for RA. An association with smoking and RA is found particularly for rheumatoid factor (RF)-positive RA compared with RF-negative disease [48,49]. Current smokers or ex-smokers have the potential to develop autoantibody-positive RA, with an odds ratio of 1.7 to 1.9. This risk increases with cumulative smoking dose [48]. A recent report investigated whether smoking is associated primarily with the development of RF or anti-CCP antibodies. This study revealed a gene–environment interaction by showing that, in the presence of HLA-shared epitope alleles, smoking contributes significantly to the development of anti-CCP antibodies [50].

A predictive effect of oral anticonceptives on RA has been claimed [51]. This finding was, however, not replicated in the Nurses' Health study [52].

Predictive value of a DNA fingerprint test

A large problem in transferring the data on genetic risk factors to prediction models is that the most current studies compared patients with controls, revealing odds ratios that are determined on group levels. The value of these genetic risk factors for individual predictive testing may be limited. Compare, for example, the statistical models to predict the pre-test probability of *BRCA1/2* genes. *BRCA1* or *2* carriers have a very strong risk for ovarian/breast cancer. The statistical models to predict the presence of a *BRCA1/2* risk allele are only informative in a selected population with affected family members [53]. This example underlines that findings for a whole group cannot be used automatically for prediction in subgroups of patients or for individuals. Genes may confer risk to subgroups of RA patients. For example, the well known HLA shared-epitope alleles particularly predispose an individual to anti-CCP-positive RA [17]. The *BRCA* example also elucidates that the predictive value of a test depends on the prevalence of a disease in a population. For UA, a number of inception cohorts of patients with recent onset arthritis have identified patients with a form of arthritis that has the potential for a persistent course, without fulfilling the classification criteria of other rheumatic disorders [54]. In nine cohorts, the proportion of patients with UA that evolved into RA within 1 year varied from 17% to 32%. Thus, in this group of UA patients, the pre-test probability of developing RA varies between 17% and 32%. Given the dynamics of UA development to either remission or progression to RA, the evaluation of predictive models for this patient group is highly relevant.

Characteristics of the disease process & prediction

The theoretical background of this section is the assumption that the expression of the disease in an initial phase allows prediction of the outcome. The genomic revolution has fuelled optimism that gene expression profiles allow such outcome measures. Gene expression profiles are used currently in breast cancer to select the patients that would benefit from adjuvant therapy [55]. However, others warned that the prognostic value of the published microarray results in cancer studies should be considered with caution, as the list of genes identified as predictors of prognosis was highly unstable and the molecular signatures depended strongly on the selection of patients [56]. The prognostic value of the microarrays used in oncology therefore needs replication.

Disease characteristics occuring at the presentation of UA that predict progression to RA

The most important and best-validated disease characteristics with regard to prediction are auto-antibodies (anti-CCP and RF) and the presence of erosions on the radiographs of hands and feet at initial presentation. In univariate analysis, the presence of anti-CCP antibodies in patients with UA conferred an odds ratio of 38 to develop RA compared with anti-CCP-negative patients with UA [57]. A logistic regression model showed an odds ratio of 16 for anti-CCP antibodies in the prediction of RA [58]. Raza and colleagues followed 124 patients who had had synovitis for less than 3 months for 72 weeks and assessed the prognostic value of anti-CCP antibodies and RF [59]. In this study, the combination of anti-CCP antibodies and RF had a positive predictive value of 100% and a negative predictive value of 88% for an RA diagnosis [59].

Clinical disease characteristics of 329 UA patients that presented to the Leiden Early Arthritis Clinic differed among those who developed RA versus those who did not. Disease characteristics associated with RA development were:

- Higher age (55 vs 46 years)
- Female sex (66 vs 50%)
- Duration of morning stiffness (60 min vs 15 min)
- Longer duration of symptoms (131 vs 81 days)
- A higher number of swollen joints (4 vs 2) [1]

Visser and colleagues developed a clinical model for the prediction of three forms of arthritis outcome: self-limiting disease, persistent nonerosive disease and persistent erosive disease [60]. For the development of this model, the first 524 consecutive patients referred to the Leiden Early Arthritis Clinic were studied and arthritis outcome was recorded after 2 years of follow-up. This prediction model consisted of seven variables:

- Symptom duration at first visit
- Morning stiffness for greater than 1 h
- Arthritis in more than 3 joints
- Bilateral compression pain in the metatarsophalangeal joints
- RF positivity
- Anti-CCP positivity
- Presence of erosions at study entry

The receiver-operating characteristic (ROC) area under the curve for discrimination between self-limiting and persistent nonerosive arthritis was 0.84 and for discrimination between persistent nonerosive and erosive arthritis it was 0.91 [60]. The addition of predisposing HLA class II alleles did not improve the discriminative ability of the model significantly [60]. The last finding might indicate that, for RA diagnostics, clinical parameters are stronger predictors than genetic markers. The model derived by Visser used all patients of the Leiden Early Arthritis Clinic, rather than only the UA patients. The advantage of the Visser model is that it can be used for a 'random' patient with arthritis who visits a rheumatological outpatient clinic. The disadvantage is that it also predicts occurrence of RA in patients who already fulfill the classification criteria for RA. Whether or not the clinical characteristics used in this model also have predictive value in patients with UA is under analysis currently.

To what extent can clinical observation be used in prediction?

Clinical observation of the natural course is the best way of predicting what the subsequent course will be. From a retrospective viewpoint, the history of the patients can be used as illustrated in the model proposed by Visser, in which a long duration of complaints was associated with higher odds for chronic and erosive disease [60]. The decision to include a 'wait and see' policy can only be taken when the possible disadvantages are also considered. The progress from UA to RA is characterized by the acquisition of certain phenotypic characteristics that form the ACR classification criteria, including joint destruction with subsequent deformities and extra-articular features, such as nodules. Given the accumulating evidence that appropriate therapy might prevent the development of a detrimental RA phenotype, observation without treatment is, in the authors' view, only justified when the patient does not fulfill the ACR criteria.

Specific studies that compare the initiation of treatment as a function of disease duration are scarce. However, valuable data were obtained in a 5-year follow-up study by Egsmose and colleagues, in which early treatment with intramuscular gold was compared with a delayed treatment strategy [61]. The early treatment group showed improvement with respect to signs and symptoms, physical function and radiographic progression, thus supporting the hypothesis of a therapeutic window of opportunity. In another trial by van der Heide and colleagues, immediate versus delayed introduction of disease-modifying anti-rheumatic drug (DMARD) therapy were compared in patients with RA diagnosed recently [62]. Early introduction of DMARDs showed greater patient improvement with regards to signs and symptoms, physical function and radiographic progression. In an observational study, Van Aken and colleagues compared the conventional pyramid strategy, consisting of sequential use of nonsteroidal anti-inflammatory drugs (NSAIDs) and subsequent DMARD therapy with immediate initiation of DMARD therapy [63]. Again, the early treatment group showed less radiographic progression [63]. Finally, an observational study performed at the Norfolk Arthritis Register provided evidence that patients in whom DMARD therapy was initiated within 6 months of RA diagnosis had a better 5-year radiographic outcome than

patients starting DMARD therapy 6 months after RA diagnosis [64]. All the aforementioned studies have investigated the importance of treatment timing with regard to diagnosis.

In summary, a role for clinical observation in the prediction of RA development seems only justified in UA patients with a low probability of developing RA.

Conclusion

In the search for methods to predict RA accurately, current data indicate that host characteristics are relevant. The identification of these host characteristics has yielded HLA alleles as being both a risk and protective and identified *PTPN22* as the second risk gene. Progress to identify the genetic risk factors that determine the remainder of the risk is slow. This is owing to the fact that each gene probably has a very small effect and to a lack of good statistical models to analyze combinations of genetic risk factors. Clinical factors that should be included in a predictive model are probably the duration of morning stiffness, the presence of an anti-CCP response and the presence of erosive abnormalities on x-rays of hands and feet.

Future perspective

Prediction of the future is impossible, but a model can provide a probability for an individual patient. Such a prediction model should guarantee a clinician and patient enough certainty (e.g., 80%) that a patient is assigned

Executive summary

Basic items in clinical prediction models

- Variability of the host.
- Variability of the causative agent.
- Variability in the disease process.
- Dynamics of the interaction that takes time into account.

Variability of incidence rates in rheumatoid arthritis

- Migration studies have similar results and host factors determine differences in rheumatoid arthritis (RA) incidence rates.
- Worldwide occurrence indicates that factors occurring during life commonly trigger RA.
- The frequency of the known genetic risk factor, human leukocyte antigen (HLA) alleles, in the different populations correlates with RA frequency.
- Absence of RA in rare, isolated populations disappears after inter-racial marriages.

Genetic risk factors for rheumatoid arthiritis

- The total contribution of genetic factors to disease susceptibility is approximately 50–60%.
- Currently identified and widely replicated genetic risk factors are *HLA* and protein tyrosine phosphatase nonreceptor (PTPN) 22.
- These factors explain approximately one third of the total genetic risk, thus the other two thirds are yet to be identified.

Environmental risk factors for rheumatoid arthritis

- Smoking is a risk factor for autoantibody-positive RA.

Predictive value of a DNA fingerprint test

- The predictive value of a test depends on the disease prevalence in the population for which the test is evaluated.
- Cohorts of patients with recent onset arthritis referred to a rheumatologist revealed a significant proportion of patients with undifferentiated arthritis (UA).
- Approximately one third of patients with UA develop RA.
- The group of patients with UA is the most relevant group for evaluating predictive models.

Disease characteristics that determine whether undifferentiated arthritis progresses to rheumatoid arthritis

- Presence of anti-CCP antibodies (or rheumatoid factor).
- Joint destruction on x-rays of hands and feet.
- Clinical characteristics, such as the presence of morning stiffness, higher number of swollen joints and bilateral compression pain of metatarsophalangeal joints.

Relevance of clinical observation in prediction models for rheumatoid arthiritis

- Long symptom duration is associated with a higher chance of developing RA.
- Early disease-modifying anti-rheumatic drug treatment prevents acquisition of a detrimental phenotype in patients that fill American College of Rheumatology RA criteria, thereby limiting the place of 'wait and see' strategies.

to the correct category. In the context of UA, where approximately one third will develop RA and two thirds will not, it is not known exactly what the minimum value of the the fraction of explained variation (R^2) has to be to result in a valuable prediction model. The R^2 is a measure of the model's ability to predict. It compares the mean squared error of the prognostic model with the mean squared error of the model without any prognostic variables and does not have a dimension. Some indication of an acceptable R^2 can be obtained from a similar problem, the prediction of the severity of joint destruction in RA. Recently, De Vries and colleagues determined the adequacy of clinical parameters in the prediction of joint destruction [Unpublished Data]. This model had an R^2 of 0.64 and classified 62% of patients correctly. Furthermore, it was calculated that to classify of 80% of the patients correctly, such a hypothetical model should have a R^2 of 0.9 [Unpublished Data]. A model that predicts joint damage scores gives an estimate for a continuous variable, and is therefore different from a model that predicts the absence or presence of RA development. Nevertheless, the data as presented by De Vries and colleagues indicate the requirements for a model to predict disease development adequately in patients with UA. To our knowledge, there are currently no prediction models analyzed that are able to determine with at least 80% certainty whether an individual patient will develop RA or not. However, given that more genetic factors associated with RA susceptibility will probably be identified in the next decade, we expect that these results will be included in future prediction models.

The predictive value of disease characteristics, such as anti-CCP antibodies, has already been identified and, given its large and specific effect, this will be included in prediction models. Clinical characteristics have not yet been defined in great detail, but we expect that, with the current inclusion of many patients in different early arthritis initiatives, these data will become available in the next decade. Given the expectation that genetic, serological and clinical data each contain independent information, it should be possible to combine these data sets to gain more accurate prognostic information. Hopefully the R^2 of such a model will be large enough to allow prediction at the patient level.

This review has focused on prediction of the diagnosis of RA, but not the prognosis of RA. This is because of the lack of epidemiological data regarding whether RA severity, such as rate of joint destruction, is caused by, for example, genetic factors. In the next decade, we expect that these basic epidemiological data will become available, therefore leading onto the development of predictive tests for RA outcome.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. van Aken J, van Dongen H, le Cessie S, Allaart CF, Breedveld FC, Huizinga TWJ: Comparison of long term outcome of patients with rheumatoid arthritis presenting with undifferentiated arthritis or with rheumatoid arthritis: an observational cohort study. *Ann. Rheum. Dis.* 65, 20–25 (2006).
2. Linn-Rasker SP, Allaart CF, Kloppenburg M, Breedveld FC, Huizinga TWJ: Sustained remission in a cohort of patients with RA: association with absence of IgM-rheumatoid factor and absence of antiCCP antibodies. *Int. J. Adv. Rheumatology.* 2(4), 4–6 (2004).
3. van der Helm-van Mil AH, Dieude P, Schonkeren JJ, Cornelis F, Huizinga TW: No association between tumour necrosis factor receptor type 2 gene polymorphism and rheumatoid arthritis severity: a comparison of the extremes of phenotypes. *Rheumatology* 43(10), 1232–1234 (2004).
4. Roberts-Thomson PJ, Jones ME, Walker JG, Macfarlane JG, Smith MD, Ahern MJ: Stochastic processes in the causation of rheumatic disease. *J. Rheumatol.* 29(12), 2628–2634 (2002).
5. del Puente A, Knowler WC, Pettitt DJ, Bennett PH: High incidence and prevalence of rheumatoid arthritis in Pima Indians. *Am. J. Epidemiol.* 129(6), 1170–1178 (1989).
6. Silman AJ, Pearson JE: Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res.* 4, S265–S272 (2002).
7. Ferucci ED, Templin DW, Lanier AP: Rheumatoid arthritis in American Indians and Alaska natives: a review of the literature. *Semin. Arthritis Rheum.* 34(4), 662–667 (2005).
8. Silman A, Harrison B, Barrett E, Symmons D: The existence of geographical clusters of cases of inflammatory polyarthritis in a primary care based register. *Ann. Rheum. Dis.* 59, 152–154 (2000).
9. Roberts-Thomson RA, Roberts-Thomson PJ: Rheumatic disease and the Australian aborigine. *Ann. Rheum. Dis.* 58, 266–270 (1999).
10. Silman AJ, Ollier W, Holligan S *et al.*: Absence of rheumatoid arthritis in a rural Nigerian population. *J. Rheumatol.* 20, 618–622 (1993).
11. Doran MF, Crowson CS, O'Fallon WM, Hunder GG, Gabriel SE: Trends in incidence and mortality in rheumatoid arthritis in Rochester, Minnesota, over a forty-year period. *Arthritis Rheum.* 46(3), 625–631 (2002).
12. Shichikawa K, Inoue K, Hirota S *et al.*: Changes in the incidence and prevalence of rheumatoid arthritis in Kamitonda, Wakayama, Japan, 1965–1996. *Ann. Rheum. Dis.* 58(12), 751–756 (1999).
13. Huizinga TWJ, Linn-Rasker S, Lard RG, Westendorp RG: Genetic drift as an explanation for the reduced incidence of rheumatoid arthritis. *Arthritis Rheum.* 46(11), 3107 (2002).

14. Cummins J: Evolutionary forces behind human infertility. *Nature* 397, 557–558 (1993).
15. Uhlig T, Kvien TK: Is rheumatoid arthritis disappearing? *Ann. Rheum. Dis.* 64(1), 7–10 (2005).
16. Welsing PMJ, Fransen J, van Riels PLCM: Is the disease course of rheumatoid arthritis becoming milder? Time trends since 1985 in an inception cohort of early rheumatoid arthritis. *Arthritis Rheum.* 52(9), 2616–2624 (2005).
17. Huizinga TWJ, Amos CI, van der Helm-van Mil AH *et al.*: Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum.* In press (2005).
•• Showing that positive and negative anticyclic citrullinated peptide (CCP) rheumatoid arthritis (RA) patients have different genetic risk factors.
18. Rantapaa-Dahlqvist S, de Jong BA, Berglin E *et al.*: Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis Rheum.* 48, 2741–2749 (2003).
19. Nielen MM, van Schaardenburg D, Reesink HW *et al.*: Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. *Arthritis Rheum.* 50, 380–386 (2004).
20. van der Helm-van Mil AH, Verpoort KN, Breedveld FC *et al.*: Antibodies to citrullinated proteins and differences in clinical progression of rheumatoid arthritis *Arthritis Res. Ther.* In press (2005).
21. MacGregor A, Ollier W, Thomson W *et al.*: HLA-DRB1*0401/0404 genotype and rheumatoid arthritis: increased association in men, young age at onset, and disease severity. *J. Rheumatol.* 22, 1032–1036 (1995).
22. van der Helm-van Mil AH, Huizinga TW, Schreuder GM *et al.*: An independent role for protective HLA Class II alleles in rheumatoid arthritis severity and susceptibility. *Arthritis Rheum.* 52(9), 2637–2644 (2005).
23. Begovich AB, Carlton VE, Honigberg LA *et al.*: A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* 75(2), 330–337 (2004).
•• Nice example of a genome-wide association study using single nucleotide polymorphisms (SNPs). The excellent methodology used multiple rounds of replications to diminish the chance of false-positive results. A new genetic risk factor for RA was identified.
24. van Oene M, Wintle RF, Liu X *et al.*: Association of the lymphoid tyrosine phosphatase R620W variant with rheumatoid arthritis, but not Crohn's disease, in Canadian populations. *Arthritis Rheum.* 52(7), 1993–1998 (2005).
25. Simkins HM, Merriman ME, Highton J *et al.*: Association of the *PTPN22* locus with rheumatoid arthritis in a New Zealand Caucasian cohort. *Arthritis Rheum.* 52(7), 2222–2225 (2005).
26. Wesoly J, Chokkalingam AP, van der Helm-van Mil AH *et al.*: Association of the PTPN22 C1858T SNP with RA phenotypes in an inception cohort. *Arthritis Rheum.* In press (2005).
27. Bottini N, Musumeci L, Alonso A *et al.*: A functional variant of lymphoid tyrosine phosphatase is associated with Type I diabetes. *Nat. Genet.* 36(4), 337–338 (2004).
28. Criswell LA, Pfeiffer KA, Lum RF *et al.*: Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the *PTPN22 620W* allele associates with multiple autoimmune phenotypes. *Am. J. Hum. Genet.* 76(4), 561–571 (2005).
29. Swanberg M, Lidman O, Padyukov L *et al.*: MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. *Nat. Genet.* 37(5), 486–494 (2005).
30. Suzuki A, Yamada R, Chang X *et al.*: Functional haplotypes of *PADI4*, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* 34(4), 395–402 (2003).
31. Caponi L, Petit-Teixeira E, Sebbag M *et al.*: A family-based study shows no association between rheumatoid arthritis and the *PADI4* gene in a French Caucasian population. *Ann. Rheum. Dis.* 64(4), 587–593 (2004).
32. Barton A, Bowes J, Eyre S *et al.*: A functional haplotype of the *PADI4* gene associated with rheumatoid arthritis in a Japanese population is not associated in a United Kingdom population. *Arthritis Rheum.* 50(4), 1117–1121 (2004).
33. Tokuhiro S, Yamada R, Chang X *et al.*: An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nature Genet.* 35(4), 341–348 (2003).
34. Wesoly JZ, Toes REM, Slagboom PE, Huizinga TWJ: RUNX1 intronic SNP is not associated with rheumatoid arthritis susceptibility in Dutch Caucasians. *Rheumatology* [Epub ahead of print] (2005).
35. Kochi Y, Yamada R, Suzuki A *et al.*: A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nature Genet.* 37(5), 478–485 (2005).
36. Wu H, Khanna D, Park G *et al.*: Interaction between RANKL and HLA-DRB1 genotypes may contribute to younger age at onset of seropositive rheumatoid arthritis in an inception cohort. *Arthritis Rheum.* 50(10), 3093–3103 (2004).
37. Lee EB, Kim JY, Kim EH *et al.*: Intercellular adhesion molecule-1 polymorphisms in Korean patients with rheumatoid arthritis. *Tissue Antigens* 64, 473–477 (2004).
38. Han SW, Kim GW, Seo JS *et al.*: *VEGF* gene polymorphisms and susceptibility to rheumatoid arthritis. *Rheumatology* 43(9), 1173–1177 (2004).
39. Prokunina L, Padyukov L, Bennet A *et al.*: Association of the *PD-1.3A* allele of the *PDCD1* gene in patients with rheumatoid arthritis negative for rheumatoid factor and the shared epitope *Arthritis Rheum.* 50(6), 1770–1773 (2004).
40. Lee YH, Kim HJ, Rho YH *et al.*: Interleukin-1 receptor antagonist gene polymorphism and rheumatoid arthritis. *Rheumatol. Int.* 24(3), 133–136 (2004).
41. Mattey DL, Dawes PT, Gonzalez-Gay MA *et al.*: *HLA-DRB1* alleles encoding an aspartic acid at position 70 protect against development of rheumatoid arthritis. *J. Rheumatol.* 28(2), 232–239 (2001).
42. Wagner U, Kaltenhäuser S, Pierer M *et al.*: Prospective analysis of the impact of HLA-DR and -DQ on joint destruction in recent-onset rheumatoid arthritis. *Rheumatology* 42, 553–562 (2003).
43. Deighton CM, Walker DJ, Griffiths ID *et al.*: The contribution of HLA to rheumatoid arthritis. *Clin. Genet.* 36, 178–182 (1989).
44. Huizinga TW: Genetics in rheumatoid arthritis. *Best Pract. Res. Clin. Rheumatol.* 17(5), 703–716 (2003).
45. Osorio Y Fortea J, Bukulmez H *et al.*: Dense genome-wide linkage analysis of rheumatoid arthritis, including covariates. *Arthritis Rheum.* 50(9), 2757–2765 (2004).
46. John S, Shephard N, Liu G *et al.*: Whole-genome scan, in a complex disease, using

11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.* 75, 54–64 (2004).

47. MacGregor AJ, Snieder H, Rigby AS *et al.*: Characterising the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* 43, 31–37 (2000).
- Gives an estimate of the genetic contribution to RA.
48. Stolt P, Bengtsson C, Nordmark B *et al.*: EIRA study group: Quantification of the influence of cigarette smoking on rheumatoid arthritis: results from a population based case–control study, using incident cases. *Ann. Rheum. Dis.* 62(9), 835–841 (2003).
49. Padyukov L, Silva C, Stolt P, Alfredsson L, Klareskog L: A gene–environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. *Arthritis Rheum.* 50(10), 3085–3092 (2004).
- The first demonstration of an interaction between an environmental risk factor and a genetic risk factor for RA.
50. Linn-Rasker SP, van der Helm-van Mil AH, van Gaalen FA *et al.*: Smoking is a risk factor for antiCCP antibodies only in RA patients that carry HLA-DRB1 shared epitope alleles. *Ann. Rheum. Dis.* [Epub ahead of print] (2005).
51. Doran MF, Crowson CS, O'Fallon WM, Gabriel SE: The effect of oral contraceptives and estrogen replacement therapy on the risk of rheumatoid arthritis: a population based study. *J. Rheumatol.* 31(2), 207–213 (2004).
52. Karlson EW, Mandl LA, Hankinson SE, Grodstein F: Do breast-feeding and other reproductive factors influence future risk of rheumatoid arthritis? Results from the Nurses' Health Study. *Arthritis Rheum.* 50(11), 3458–3467 (2004).
53. de la Hoya M, Diez O, Perez-Segura P *et al.*: Pre-test prediction models of *BRCA1* or *BRCA2* mutation in breast/ovarian families attending familial cancer clinics. *J. Med. Genet.* 40(7), 503–510 (2003).
54. Verpoort KN, van Dongen H, Allaart CF, Toes RE, Breedveld FC, Huizinga TW: Undifferentiated arthritis-disease course assessed in several inception cohorts. *Clin. Exp. Rheumatol.* 22(5 Suppl. 35), S12–S17 (2004).
55. van 't Veer LJ, Dai H, van de Vijver MJ *et al.*: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536 (2002).
56. Michiels S, Koscielny S, Hill C: Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365(9458), 488–492 (2005).
57. van Gaalen FA, Linn-Rasker SP, van Venrooij WJ *et al.*: Autoantibodies to cyclic citrullinated peptides predict progression to rheumatoid arthritis in patients with undifferentiated arthritis: a prospective cohort study. *Arthritis Rheum.* 50(3), 709–715 (2004).
58. Berglin E, Padyukov L, Sundin U *et al.*: A combination of autoantibodies to cyclic citrullinated peptide (CCP) and HLA-DRB1 locus antigens is strongly associated with future onset of rheumatoid arthritis. *Arthritis Res. Ther.* 6(4), R303–R308 (2004).
59. Raza K, Breese M, Nightingale P *et al.*: Predictive value of antibodies to cyclic citrullinated peptide in patients with very early inflammatory arthritis. *J. Rheumatol.* 32(2), 231–238 (2005).
60. Visser H, le Cessie S, Vos K *et al.*: How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. *Arthritis Rheum.* 46(2), 357–365 (2002).
- Nice study showing a model that predicts disease outcome using clinical parameters (disease persistency and erosiveness).
61. Egsmose C, Lund B, Borg G *et al.*: Patients with rheumatoid arthritis benefit from early 2nd line therapy: 5 year follow-up of a prospective double-blind placebo-controlled study. *J. Rheumatol.* 22, 2208–2213 (1995).
62. van der Heide A, Jacobs JW, Bijlsma JW *et al.*: The effectiveness of early treatment with 'second-line' antirheumatic drugs. A randomized, controlled trial. *Ann. Intern. Med.* 124, 699–707 (1996).
63. van Aken J, Lard LR, le Cessie S *et al.*: Radiological outcome after four years of early versus delayed treatment strategy in patients with recent onset rheumatoid arthritis. *Ann. Rheum. Dis.* 63, 274–279 (2004).
64. Bukhari MA, Wiles NJ, Lunt M *et al.*: Influence of disease-modifying therapy on radiographic outcome in inflammatory polyarthritis at five years: results from a large observational inception study. *Arthritis Rheum.* 48, 46–53 (2003).

Affiliations

- *Annette HM van der Helm-van Mil, MD*
  Leiden University Medical Center, Department of Rheumatology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands
  Tel.: +31 715 263 598;
  Fax: +31 715 266 752;
  AvdHelm@lumc.nl
- *Tom WJ Huizinga*
  Leiden University Medical Center, Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands