**TITLE PAGE**

**Full title:** Development of predictive genetic tests for improving the safety of new medicines: a concept paper on the utilisation of routinely collected electronic health records

**Short title:** Predictive genetic test development using electronic health records

**Authors:**
Kevin Wing[1]
Ian Douglas[1]
Krishnan Bhaskaran[1]
Olaf H. Klungel[2] *
Robert F. Reynolds[3]
Munir Pirmohamed[4]
Liam Smeeth[1]
Tjeerd P. van Staa[1, 2, 5]

[1]Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
[2]Department of Pharmacoepidemiology, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, The Netherlands
[3]Epidemiology, Pfizer, New York, NY, USA
[4]The Wolfson Centre for Personalised Medicine, Department of Molecular and Clinical Pharmacology, University of Liverpool, UK
[5]Clinical Practice Research Datalink (CPRD), Medicines and Healthcare products Regulatory Agency, London, UK

*On behalf of the members of work-package 2 (WP2) of PROTECT (Framework for pharmacoepidemiology studies): Y. Alvarez, G. Candore, J. Durand, J. Slattery (European Medicines Agency); J. Hasford, M. Rottenkolber (Ludwig-Maximilians-Universität-München); S. Schmiedl (Witten University); F. de Abajo Iglesias, M. Gil, C. Huerta Alvarez, , E. Martin, G. Requena (Agencia Espanola de Medicamentos y Productos Sanitarios); R. Brauer, G. Downey, M. Feudjo-Tepie, Justyna Amelio (Amgen NV); S. Johansson (AstraZeneca); J. Robinson, M. Schuerch, I. Tatt (Roche); L.A. Garcia, A. Ruigomez (Fundación Centro Español de Investigación Farmacoepidemiológica); J. Campbell, A. Gallagher, E. Ng, T. Van Staa (General Practice Research Database); O. Demol (Genzyme); K. Davis, J. Logie, J. Pimenta, D. Webbs (GlaxoSmithKline Research and Development LTD); L. Bensouda-Grimaldi, R. Beau-Lejdstrom (L.A. Sante Epidemiologie Evaluation Recherche); U. Hesse, (Lægemiddelstyrelsen (Danish Medicines Agency) ); M. Miret (Merck KGaA ); J. Fortuny, P. Primatesta, E. Rivero, R. Schlienger (Novartis); A. Bate, N. Gatto, R. Reynolds (Pfizer); E. Ballarin, P. Ferrer, M. Sabate, L. Ibañez, J.R. Laporte, M. Sabaté (Fundació Institut Català de Farmacologia); V. Abbing-Karahagopian, S. Ali, A. Afonso, D. de Bakker, S. Belitser, A. De Boer, M.L. de Bruin, A.C.G. Egberts, , L. van Dijk, H. Gardarsdottir, R.H. Groenwold, M. De Groot, A.W. Hoes, O.H. Klungel, H.G.M.. Leufkens, , K.C.B. Roes, P. Souverein, F. Rutten, J. Uddin, H.A.. Van den Ham, F. de Vries (Universiteit Utrecht).

**Keywords:** adverse drug reactions, predictive genetic testing, electronic health records, genetic association studies, randomized controlled trials.

**Teaser sentence:** How major challenges associated with the development of predictive genetic tests for drug safety could be overcome using databases of routinely collected health information.

**Word count:** 2000 (not including abstract, refs, figure/section headings and figure footnotes)
Abstract word count: 120

**ABSTRACT**

Serious adverse drug reactions are an important cause of hospitalisation and can result in the withdrawal of licensed drugs. Genetic variation has been shown to influence adverse drug reaction susceptibility, and predictive genetic tests have been developed for a limited number of adverse drug reactions. The identification of people with adverse drug reactions, obtaining samples for genetic analysis and rigorous evaluation of clinical test effectiveness represent significant challenges to predictive genetic test development. Using the example of serious drug-induced liver injury, we illustrate how a database of routinely collected electronic health records could be used to overcome these barriers by (1) facilitating rapid recruitment to genome-wide association studies and (2) supporting efficient randomized controlled trials of predictive genetic test effectiveness.

**MAIN TEXT**

## Introduction

Adverse drug reactions are estimated to be responsible for over 5% of hospital admissions [1,2]. Approximately 150 drugs have been withdrawn from the market since 1960 due to safety issues [3], in some cases many years after initial approval [4]. Evidence is increasing for a genetic predisposition to a number of serious adverse drug events [5]. For a limited number of drugs, observational genetic studies of association and subsequent randomised controlled trials (RCTs) of effectiveness of genotype guided treatment have enabled predictive genetic tests to be developed that have allowed valuable medicines to remain on the market with greatly improved risk benefit profiles [6].

Serious drug induced liver injury is a leading cause of drug withdrawals [3]. Associations between specific genes and susceptibility to drug-induced liver injury caused by a number of drug therapies have been identified, although predictive genetic test development has been minimal and many genes conferring susceptibility are yet to be identified [7,8]. A major challenge is the time and cost associated with finding patients who have suffered a reaction of interest, and recruiting sufficient numbers for initial genome-wide association studies and subsequent replication studies [9]. Following predictive genetic test development, an RCT to evaluate effectiveness of the genotype guided treatment versus standard care may be required [10], introducing further logistical challenges. The possibility of using databases of routinely collected electronic health records (EHRs) to support pharmacogenomics has been discussed elsewhere (Yasmina et al, unpublished and [11-13]). This paper provides further detail by illustrating how an EHR database could be used to (1) identify people who have experienced serious adverse reactions linked to a newly licensed drug in order to invite them to provide genetic samples for genome-wide association studies and (2) test the efficacy of any developed genetic test in a cluster RCT. We illustrate these ideas using drug-induced liver injury as an example adverse drug reaction.

## Identification of drug-induced liver injury within routinely collected electronic health records

Routinely collected EHRs provide the potential for low-cost, efficient epidemiological cohort identification and analysis [14]. Although database-specific, an EHR for an individual patient typically includes a unique patient id, clinical diagnoses (as standardised diagnostic codes), drug prescriptions, laboratory test results, and in some cases lifestyle information (such as smoking, drinking, BMI). Coverage of the underlying population is likely to be broad, and new linkages between databases are enhancing the ability to ascertain disease status. For example, the UK Clinical Practice Research Datalink (CPRD) primary care database contains anonymised health information for approximately 8% of the total UK population and can be linked to the UK Hospital Episode Statistics database [15].

Effective case identification algorithms can be developed that utilise EHR databases to identify cases of drug-induced liver injury. Work has been performed demonstrating the portability of such algorithms across different institutions [16] and work is ongoing to facilitate standardized implementation across databases in different countries (Ruigomez and Brauer for the IMI-PROTECT group [17], unpublished). An example algorithm is provided in Figure 1. Potential drug-induced liver injury cases are identified based upon specific diagnostic codes (routinely inputted by clinicians). The database is searched for liver test results within a specific time period from the diagnostic code indicating possible liver injury. Patient records with no liver test results or results not indicative of drug-induced liver injury are removed, and the remaining patients are considered to be characteristic of drug-induced liver injury. The type of liver injury can be determined as required

[13], before a set of exclusion codes is applied in order to remove individuals with other underlying (non-drug) causes for their liver symptoms. Verification of the cohort of potential cases can be performed by analysing the association between being an algorithm-selected case and having recently been prescribed drugs that are well known causes of drug-induced liver injury (such as flucloxacillin or amoxicillin-clavulanate [8]).

For potential cases with a prescription for the drug of interest within a specified risk period, a history of prescriptions and diagnoses (within a defined period) can be extracted from the database, enabling information on other potential causes of the liver injury to be obtained. Additional data on whether the patient was referred to hospital or not, and from linked databases could also be considered at this point (such as pathology databases or general hospital statistics databases). Brief questionnaires can then be sent to the responsible clinician as appropriate, in order to obtain any referral letters from liver specialists, and to ascertain whether the responsible clinician considers the events to represent drug-induced liver injury as a result of the drug in question. The totality of the health data obtained for each patient can then be reviewed by medically trained professionals and considered against international causality criteria [18], in order to identify likely cases.

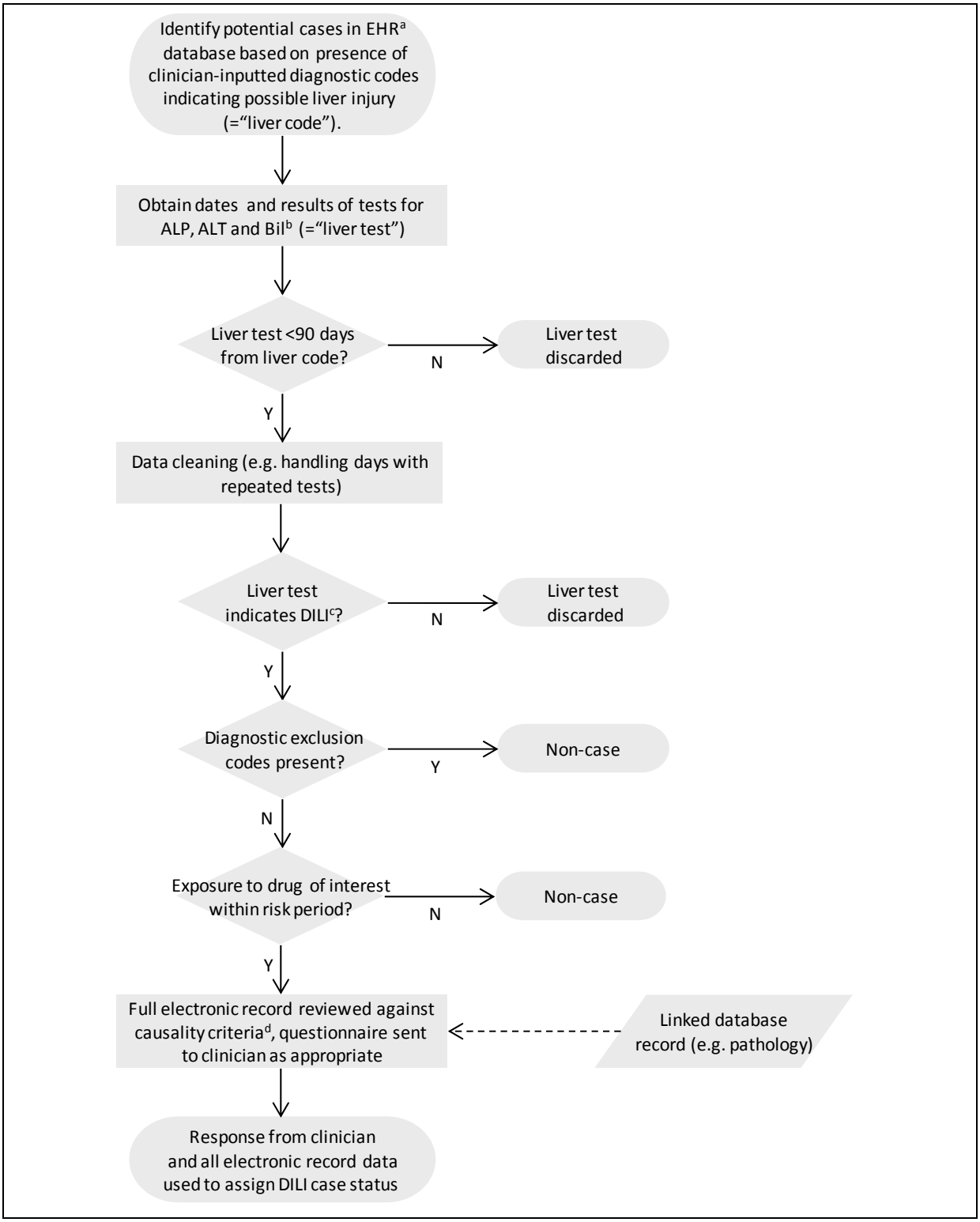


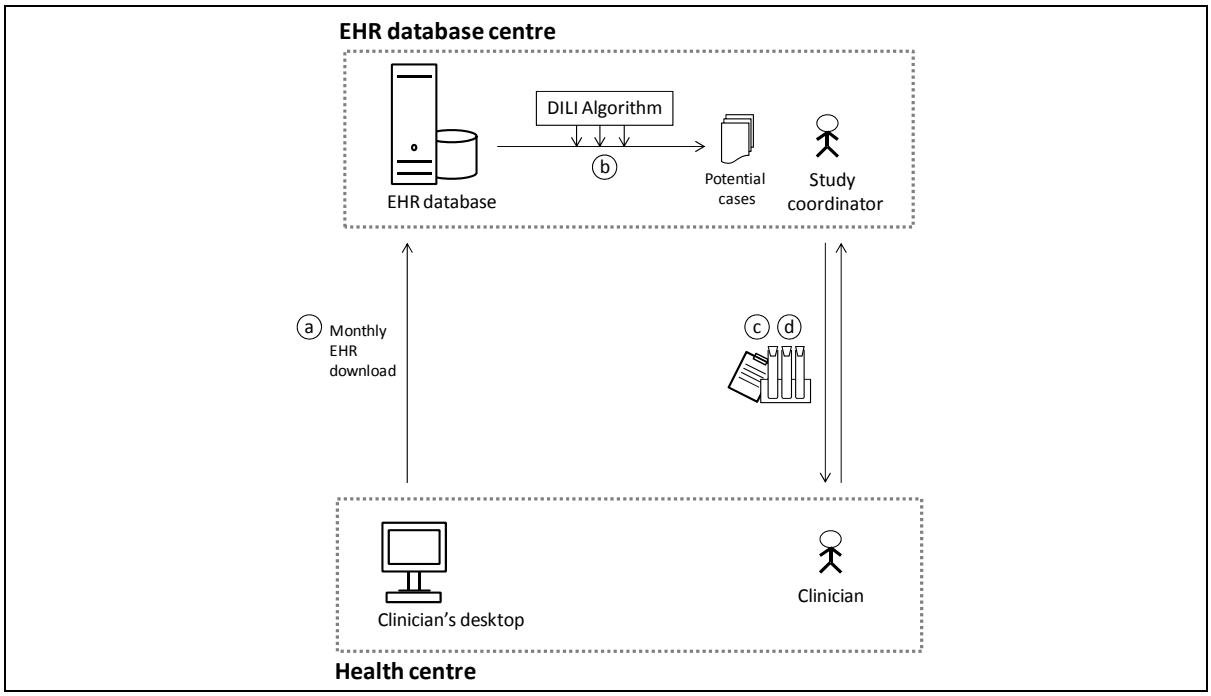

**Figure 1: Example algorithm for identifying cases of drug-induced liver injury using a database of electronic health records**

[a]EHR: electronic health record
[b]ALP, ALT and Bil**:** ALP=Alkaline phosphatase, ALT=Alanine aminotransferase, Bil=Bilirubin
[c]Drug-induced liver injury: serious Drug induced liver injury, defined as ALT>5xULN or ALP>2xULN or ALT>3xULN with Bil>2xULN [13]
[d]Causality criteria: as determined by international consensus [18]

**Active monitoring of EHRs for recruitment to genetic association studies**

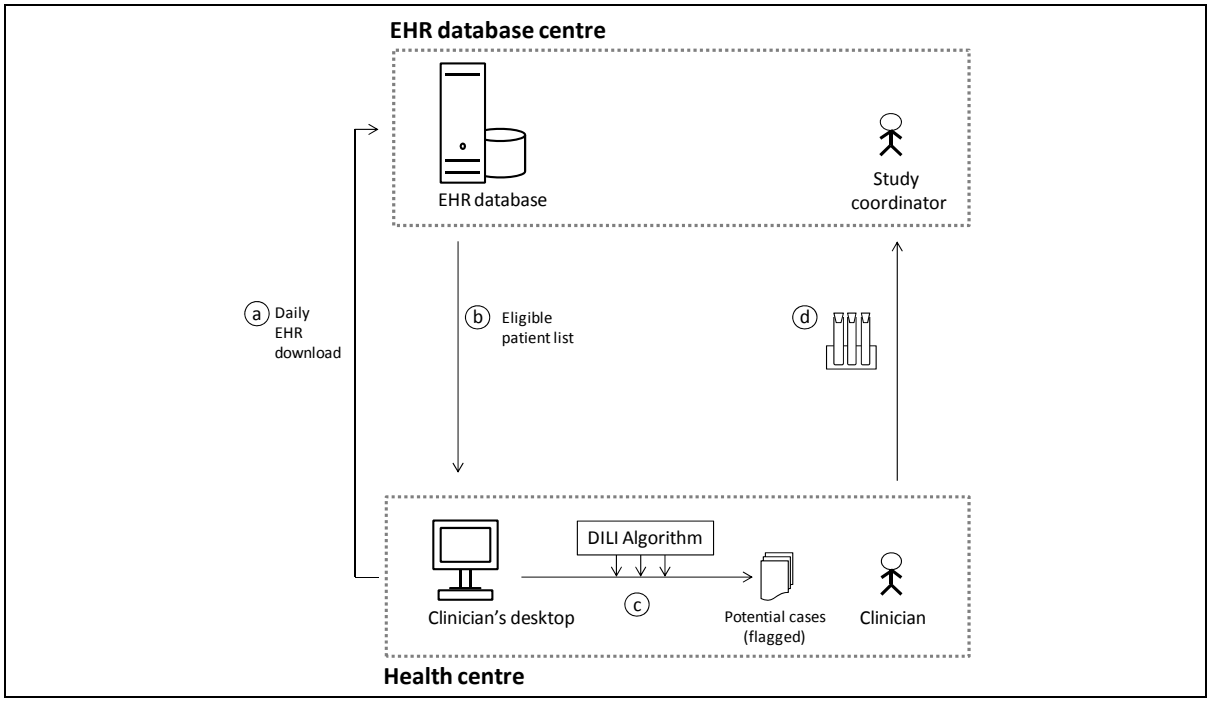
Drug-induced liver injury algorithms as described have typically been applied to database study populations at a single time-point following drug registration, in order to retrospectively identify cases for inclusion in epidemiological studies [19,20]. For recruitment to genetic association studies, we propose an alternative "active monitoring" model. Genetic sampling kits would be sent directly to the clinician of individuals identified as cases by continuous database surveillance, and the clinician would obtain consent and take a blood sample to be sent to the study coordinator.

Two possible information processing approaches could support this model. The first approach would utilise the regular (e.g. monthly) download of EHRs from contributing health centres to the central database. The drug-induced liver injury algorithm could be run against the database following each download, and the clinicians of selected potential cases could then be invited to complete the short questionnaire described previously, before a genetic sampling kit is sent to them if appropriate (Figure 1). The second approach would be to utilise technology comparable to that installed on clinician desktops for performing pragmatic randomized trials within the UK CPRD database [21]: patients prescribed the drug of interest would be identified at the EHR database centre and added to lists of eligible patients. The algorithm would be run locally on the clinician's desktop when an eligible patient had their health record updated, and if potential drug-induced liver injury is suggested, the patient would be flagged as an eligible potential case. The clinician would make the final decision on case status, and obtain and provide genetic samples using pre-provided genetic sampling kits (Figure 2).




(a) Updated EHRs are uploaded from the clinician's desktop to the EHR database according to the usual protocol (e.g. monthly) (b) The drug-induced liver injury algorithm is applied to the database of EHRs, in order to identify potential cases (c) The study coordinator sends a short questionnaire to the responsible clinician of all the potential cases, and the coordinator uses this response and all the electronic record information to identify likely cases (d) The study coordinator sends a genetic sampling kit to the clinician, who obtains consent from the likely case before taking genetic samples and sending back to the study coordinator

**Figure 2: Actively monitoring an EHR database in order to recruit to drug-induced liver injury genetic association studies – approach one**




(a) Daily downloads of information on drug prescriptions are transferred between the clinician's desktop and the EHR database (b) This updates an eligible patient list within the EHR database, which is sent to the clinician's desktop each time it is updated (c) The algorithm is run on the clinician's desktop for eligible patients (d) Potential cases are flagged to the clinician who makes a decision on case status (e) Genetic samples for likely cases are provided by the clinician to the study coordinator

**Figure 3: Actively monitoring an EHR database in order to recruit to drug-induced liver injury genetic association studies – approach two**

**EHR-based randomised controlled trials of predictive genetic test effectiveness**

Following development of a potential predictive test, an RCT would provide rigorous evidence that genotype guided treatment improves health outcomes [10,22]. Existing EHR database infrastructure could allow a cluster RCT to be set up in which contributing health centres are randomly assigned to implement use of the predictive test or not, an approach that is already being applied for other disease areas [23]. In the intervention arm, patients with an indication for the drug of interest would be tested for the gene of susceptibility and the results could be used to inform subsequent treatment and monitoring; in non-intervention arm health centres treatment would be based on information available according to standard care only. Follow-up of the participating health centres could then be performed by using the algorithm to monitor EHR data on a monthly basis, in order to assess whether the rate of adverse drug reaction for the drug of interest differs between health centres randomised to use of the predictive test and those randomised to give standard care. An evaluation of subsequent changes in treatment approach based upon predictive test result could also be performed in this way.

**Discussion**

In an ideal world, one might investigate genetic determinants of drug-induced liver injury due to a newly licensed drug by setting up a multi-centre prospective observational study at hospital liver clinics, in which genetic samples are taken rapidly after a patient with a drug reaction is identified. However, such studies require a dedicated infrastructure and are costly to perform. Biobanks represent an alternative source of genetic information. Typically, these are repositories of prospectively stored genetic information for very large numbers of patients [24] which provide valuable resources for studying (relatively common) complex gene-disease genetic associations. Unfortunately, the typically low frequency of SAEs means that recruiting sufficient patients from biobanks to carry out an adequately-powered GWAS study within a short timeframe following drug registration would be difficult [25]. Further disadvantages include the high financial and environmental cost of each repository, and the possibility that a lack of inter-biobank standardisation could result in measurement error within studies across biobanks [26].

One possible approach for addressing these challenges is the formation of collaborative groups such as the International Serious Adverse Event Consortium (iSAEC),a large-scale private-public biomedical consortium [27] [28]. Genetic samples obtained from large prospective studies and biobanks are shared between stakeholders, with an indication of the resource applied to this effort provided by the fact that 6 of the collaborators are top-ten ranked pharmaceutical companies [29]. This level of collaboration has contributed significantly to recent progress in the identification of adverse drug reaction gene associations [30].

Our proposed model provides an alternative to an iSAEC-type approach, could be implemented at relatively low cost, and would allow active monitoring and recruitment for newly licensed drugs. By continuous retrospective selection of recent drug-induced liver injury from a population-based database of routinely collected EHRs, the beneficial characteristics of a large prospective multi-centre prospective study are conferred, but with the most costly components already in place (i.e. multiple study centres, on-site study staff and individual patient records). Furthermore, the same infrastructure could be utilised for multiple drugs, the population being screened for the reaction is inherently very large, detection does not rely on a hospital referral, and a source of population-based controls for case-control studies of genetic association is provided. Data are also routinely collected that would allow analysis of some co-existing clinical and environmental determinants of

susceptibility to adverse drug reactions (another factor that may have contributed to a lack of progress in this area [9]). A comparable approach has enabled recruitment of over 700 patients from the UK CPRD for a statin-induced myopathy genetic association study, demonstrating its feasibility [31]. The ongoing work within the IMI-PROTECT consortium to standardize definitions across databases opens up the possibility of rapid recruitment for a single gene-association study across multiple databases (such as the UK CPRD, Dutch Mondrian and Spanish BIFAP databases, for example) [17]. As EHR databases start to become linked to biobanks [11], genetic samples collected in this way could be stored for use in future studies of the specific adverse drug reaction.

We also propose a role for EHRs in performing RCTs of genetic test effectiveness [21]. The costs of a conventional multi-centre RCT are likely to represent a serious obstacle to progress, and results may have limited generalisability [21]. Adoption of a recently proposed model for cluster RCTs performed within an EHR database [23] could address both of these challenges: whole health centres could be randomised to the use of a newly developed predictive genetic test, with minimum set-up costs, maximum reusability and a focus on test effectiveness in real-world settings.

Although this article focuses on the development of predictive genetic tests, we feel that the active monitoring approach described compares favourably with the current passive yellow card system for pharmacovigilance, and would also support adaptive licensing [32]. Spontaneous reporting systems suffer from reporting bias, missing denominator information and vague or incomplete case definitions, problems that would be minimised by applying real-time detection to one or multiple EHR databases. Targeted detection of adverse drug reactions observed during drug development could begin immediately following product launch (from Day 1), and could be included as part of iterative data gathering within an adaptive licensing framework.

## Conclusion

It is of critical importance to try and ensure that innovative products are not removed from the market because of risks associated with detectable genetic variation in the population, as has occurred previously [33]. In a climate where urgently needed new drugs (such as antibiotics) could potentially have a worldwide impact on public health, unnecessary withdrawals could have severe consequences. Databases of EHRs can facilitate the development of rapid and low-cost predictive genetic tests which could in turn prevent avoidable withdrawals, and reduce the number of people unnecessarily exposed to serious adverse drug reactions.

**Acknowledgements:** The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium, www.imi-protect.eu) which is a public-private partnership coordinated by the European Medicines Agency.

**Funding:** The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement n° 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. In the context of the IMI Joint Undertaking (IMI JU), the London School of Hygiene and Tropical Medicine and the Department of Pharmacoepidemiology, Utrecht University, received direct financial contributions from Pfizer. Additional funding sources are detailed below.

Ian Douglas is funded by an MRC methodology fellowship.

Krishnan Bhaskaran is funded by a post-doctoral fellowship from the National Institute of Health Research.
Liam Smeeth is funded by a Wellcome Trust Senior Clinical Fellowship.
Munir Pirmohamed and Liam Smeeth are NIHR Senior Investigators.

The views expressed are those of the authors only and not of their respective institutions, companies, or other bodies/committees that they may be associated with.

**REFERENCES**


1 Pirmohamed, M. et al. (2004) Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ* 329 (7456), 15-19

2 Kongkaew, C. et al. (2008) Hospital Admissions Associated with Adverse Drug Reactions: A Systematic Review of Prospective Observational Studies. *The Annals of Pharmacotherapy* 42 (7), 1017-1025

3 Zhang, W.R., M. W.; Chen W.; Fan, L.; Zhou, H. (2012) Pharmacogenetics of Drugs Withdrawn From the Market. *Pharmacogenomics* 13 (2), 223-231

4 EMA. (2010) European Medicines Agency recommends suspension of Avandia, Avandamet and Avaglim. (Office, P., ed.), EMA

5 Daly, A.K. (2010) Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics* 11, 241-246

6 Mallal, S. et al. (2008) HLA-B*5701 Screening for Hypersensitivity to Abacavir. *New England Journal of Medicine* 358 (6), 568-579

7 Alfirevic, A. and Pirmohamed, M. (2012) Predictive Genetic Testing for Drug-Induced Liver Injury: Considerations of Clinical Utility. *Clin Pharmacol Ther* 92 (3), 376-380

8 Daly, A.K. (2010) Drug-induced liver injury: past, present and future. *Pharmacogenomics* 11 (5), 607-611

9 Pirmohamed, M. (2011) Pharmacogenetics: past, present and future. *Drug Discovery Today* 16 (19–20), 852-861

10 van der Baan, F.H. et al. (2011) Pharmacogenetics in randomized controlled trials: considerations for trial design. *Pharmacogenomics* 12 (10), 1485-1492

11 Wilke, R.A. et al. (2011) The Emerging Role of Electronic Medical Records in Pharmacogenomics. *Clin Pharmacol Ther* 89 (3), 379-386

12 Kohane, I.S. (2011) Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 12 (6), 417-428

13 Aithal, G.P. et al. (2011) Case Definition and Phenotype Standardization in Drug-Induced Liver Injury. *Clin Pharmacol Ther* 89 (6), 806-815

14 Langan, S.M. et al. (2013) Setting the RECORD straight: developing a guideline for the REporting of studies Conducted using Observational Routinely collected Data. *J Clin Epidemiol* 5, 29-31

15 Bazelier, M.T. et al. (2011) The risk of fracture in patients with multiple sclerosis: The UK general practice research. *Journal of Bone and Mineral Research* 26 (9), 2271-2279

16 Overby, C.L. et al. (2013) A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. *Journal of the American Medical Informatics Association*

17 EMA. The Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) Objectives of PROTECT. (Vol. 2012)

18 Danan, G. and Benichou, C. (1993) Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: Application to drug-induced liver injuries. *Journal of Clinical Epidemiology* 46 (11), 1323-1330

19 Russmann, S. et al. (2005) Risk of cholestatic liver disease associated with flucloxacillin and flucloxacillin prescribing habits in the UK: Cohort study using data from the UK General Practice Research Database. *British Journal of Clinical Pharmacology* 60 (1), 76-82

20 De Abajo, F.J. et al. (2004) Acute and clinically relevant drug-induced liver injury: A population case-control study. *British Journal of Clinical Pharmacology* 58 (1), 71-80

21 Staa, T.-P.v. et al. (2012) Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ* 344

22 Smeeth, L. and Staa, T.v. (2012) Will the revolution in genetics improve healthcare? *BMJ* 345

23 Dregan, A. et al. (2012) Cluster randomized trial in the general practice research database: 2. Secondary prevention after first stroke (eCRT study): study protocol for a randomized controlled trial. *Trials* 13 (1), 181

**24** **McCarty, C. et al. (2011) The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics* 4 (1), 13**

**25** **Ritchie, M.D. (2012) The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. *Human Genetics* 131 (10), 1615-1626**

**26** **Swanson, J. (2009) The Changing Face of Biobanks. In *Clinical Genomics* (Vol. 2013) (Technology, G., ed.)**

**27** **iSAEC. (2012) About iSAEC - letter from the chairman. (Vol. 2013)**

**28** **Contreras, J.L. et al. (2013) The International Serious Adverse Events Consortium's data sharing model. *Nat Biotech* 31 (1), 17-19**

**29** **Roth, G.Y. (2012) Top 20 Pharma Report. In *Contract Pharma* (Vol. 2013) (Pharma, C., ed.)**

**30** **Daly, A.K. et al. (2009) HLA-B[ast]5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat Genet* 41 (7), 816-819**

**31** **Carr, D.F. et al. (2013) SLCO1B1 Genetic Variant Associated With Statin-Induced Myopathy: A Proof of Concept Study Using the Clinical Practice Research Datalink (CPRD). *Clin Pharmacol Ther***

**32** **Eichler, H.G. et al. (2012) Adaptive Licensing: Taking the Next Step in the Evolution of Drug Approval. *Clin Pharmacol Ther* 91 (3), 426-437**

**33** **Aithal, G.P. and Daly, A.K. (2010) Preempting and preventing drug-induced liver injury. *Nat Genet* 42 (8), 650-651**