ARTICLE IN PRESS

npj Breast Cancer

https://doi.org/10.1038/s41523-026-00931-2

Article in Press

A multi-ancestry genome-wide study of tamoxifen metabolism and breast cancer recurrence

Received: 3 December 2025

Accepted: 3 March 2026

Cite this article as: Khor, C.C., Ong, W.S., Lim, E.H. *et al.* A multi-ancestry genome-wide study of tamoxifen metabolism and breast cancer recurrence. *npj Breast Cancer* (2026). https://doi.org/10.1038/s41523-026-00931-2

Chiea Chuen Khor, Whee Sze Ong, Elaine Hsuen Lim, Etienne Chatelut, Sylvia Chen, Thomas E. Muerdter, Jean E. Abraham, Natalia Sutiman, Joanne Siok Liu Lim, Sze Sing Lee, Cecile Arellano, Zheng Li, Kar Seng Sim, Diana M. Eccles, William J. Tapper, Tom Maishman, Nathalie K. Zgheib, Stefan Winter, Boian Ganchev, Leila Dorling, Qi Guo, Carlos Caldas, Helena M. Earl, Louise Hiller, Janet Dunn, Raymond Chee Hui Ng, Yoon Sim Yap, Mabel Wong, Fuh Yong Wong, Nan Soon Wong, Peter Cher Siang Ang, Rebecca Dent, Peter Krippl, Uwe Langsenlehner, Tanja Langsenlehner, Arafat Tfayli, Elke Schaeffeler, Michel Eichelbaum, Ute Hamann, Peter A. Fasching, Matthias W. Beckmann, Florence Dalenc, Melanie White-Koning, Hiltrud B. Brauch, Werner Schroth, Wilfried Renner, Matthias Schwab, Fabienne Thomas & Balram Chowbay

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

A multi-ancestry genome-wide study of tamoxifen metabolism and breast cancer recurrence.

Chiea Chuen Khor[1,2] , Whee Sze Ong[3], Elaine Hsuen Lim[4], Etienne Chatelut[5,6], Sylvia Chen[7], Thomas E. Muerdter[8,9], Jean E. Abraham[10,11,12], , Natalia Sutiman[7], Joanne Siok Liu Lim[7], Sze Sing Lee[7], Cecile Arellano[5] , Zheng Li[1], Kar Seng Sim[1], Diana M Eccles[13,14], William J Tapper[13,14], Tom Maishman[13,14], Nathalie K. Zgheib[15], Stefan Winter[8,9], Boian Ganchev[8,9], Leila Dorling[16], Qi Guo[17], Carlos Caldas[10,12,18], Helena M. Earl[10,11,12], Louise Hiller[19], Janet Dunn[19], Raymond Chee Hui Ng[4], Yoon Sim Yap[4], Mabel Wong[4], Fuh Yong Wong[20], Nan Soon Wong[21], Peter Cher Siang Ang[21], Rebecca Dent[4], Peter Krippl[22], Uwe Langsenlehner[23], Tanja Langsenlehner[23], Arafat Tfayli[24], Elke Schaeffeler[8,9], Michel Eichelbaum[8,9], Ute Hamann[25], Peter A. Fasching[26], Matthias W. Beckmann[27], Florence Dalenc[5,6], Melanie White-Koning[5], Hiltrud B. Brauch[8,28,29], Werner Schroth[8,9], Wilfried Renner[30], Matthias Schwab[8,28,29,31], Fabienne Thomas[5,6], Balram Chowbay*[7,32] (corresponding author)

[1]Division of Human Genetics, Genome Institute of Singapore, Singapore, Singapore
[2]Singapore Eye Research Institute, Singapore, Singapore
[3]Division of Clinical Trials and Epidemiological Sciences, National Cancer Centre Singapore, Singapore, Singapore
[4]Division of Medical Oncology, National Cancer Centre Singapore, Singapore, Singapore
[5]Univ Toulouse, Inserm, CRCT, Toulouse, France
[6]Institut Claudius Regaud, Institut Universitaire du Cancer de Toulouse - Oncopole, Toulouse, France.
[7]Clinical Pharmacology Laboratory, Division of Cellular & Molecular Research, National Cancer Centre Singapore, Singapore, Singapore
[8]Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany
[9]University Tuebingen, Tuebingen, Germany
[10]University of Cambridge, Department of Oncology (Box 193), Addenbrooke's Hospital, Hills Road, Cambridge, United Kingdom
[11]Cambridge Experimental Cancer Medicine Centre and Cancer Research UK Cambridge Centre, Cambridge, United Kingdom
[12]Cambridge Breast Cancer Research Unit, NIHR Cambridge Biomedical Research Centre, Cambridge University Hospital NHS Foundation Trust, Cambridge, United Kingdom
[13]Faculty of Medicine, Cancer Sciences Academic Unit and University of Southampton Clinical Trials Unit, University of Southampton, Southampton, United Kingdom
[14]University Hospital Southampton NHS Foundation Trust, Southampton, United Kingdom
[15]Faculty of Medicine, Department of Pharmacology and Toxicology, American University of Beirut, Beirut, Lebanon
[16]Centre for Cancer Genetic Epidemiology, University of Cambridge, Strangeways Research Laboratory, Cambridge, United Kingdom
[17]Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Cambridge, United Kingdom
[18]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom
[19]Warwick Clinical Trials Unit, University of Warwick, Coventry, United Kingdom
[20]Division of Radiation Oncology, National Cancer Centre Singapore, Singapore, Singapore
[21]OncoCare Cancer Centre, Mount Elizabeth Novena Medical Centre, Singapore, Singapore,
[22]Department of Internal Medicine, General Hospital Furstenfeld, Furstenfeld, Austria
[23]Internal Outpatient Department, Medizinische Universitat Graz, Graz, Austria

ARTICLE IN PRESS

ARTICLE IN PRESS

[24]Hematology-Oncology Division, Department of Internal Medicine, Faculty of Medicine, American University of Beirut, Beirut, Lebanon
[25]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany
[26]University Breast Centre Franconia & University Hospital Erlangen, Erlangen, Germany
[27]Erlangen University Hospital, Department of Gynecology and Obstetrics, Comprehensive Cancer Center Erlanger-EMN, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany
[28]German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Partner Site Tübingen, Tübingen, Germany
[29]Cluster of Excellence iFIT (EXC 2180) “Image-Guided and Functionally Instructed Tumor Therapies”, University of Tübingen, Tübingen, Germany
[30]Clinical Institute of Medical and Laboratory Diagnostics, Medical University Graz, Graz, Austria
[31]Departments of Clinical Pharmacology and Biochemistry and Pharmacy, University of Tübingen, Tübingen, Germany
[32]Centre for Clinician-Scientist Development, Duke-NUS Medical School, Singapore, Singapore

ARTICLE IN PRESS

***Corresponding Author:**
Balram Chowbay, PhD
Senior Principal Investigator,
Division of Cellular & Molecular Research, Clinical Pharmacology Laboratory,
National Cancer Centre Singapore, Singapore, Singapore
30 Hospital Boulevard, Singapore 168583
Email: ctebal@nccs.com.sg
Telephone: (65) 6306 5401

ARTICLE IN PRESS

Abstract

Tamoxifen's pharmacokinetics is strongly influenced by the highly polymorphic *CYP2D6* while the influence of other genetic variants has been inconclusive. To further delineate this genotypic-phenotypic impact, we conducted a multi-ancestry genome-wide association study in 636 hormone-receptor positive (HR+) breast cancer (BC) patients treated with 20mg tamoxifen daily for ≥8 weeks and validated these genetic determinants in another 869 patients. Association with clinical outcomes was examined in 1326 non-metastatic HR+ patients receiving adjuvant tamoxifen. A genome-wide significant association with Z-endoxifen levels was observed at *CYP2D6* locus on chromosome 22 and its downstream region of *TCF20* rs932376 A>G. Both *CYP2D6* metabolizer status and *TCF20* rs932376 A>G were independent predictors of endoxifen levels in multivariable analysis. *CYP2D6* metabolizer status accounted for greater variability of mean endoxifen levels compared to *TCF20* rs932376 A>G (91.2% vs 48.8%). These findings were replicated in validation cohorts. Neither *TCF20* rs932376 nor *CYP2D6* metabolizer status was significantly associated with BC outcomes after adjustment for known prognostic factors. Our study confirmed that *CYP2D6* metabolizer status remains as the prime predictor of steady-state Z-endoxifen levels, while *TCF20 rs932376 A>G* has a smaller, independent effect. Both genetic factors were not associated with breast cancer clinical outcomes.

ARTICLE IN PRESS

ARTICLE IN PRESS

Introduction

Breast cancer is the most frequent cancer afflicting women, with its commonest subtype being hormone-responsive based on estrogen receptor (ER) and/or progesterone receptor (PR) positivity (luminal subtype). Tamoxifen is an effective adjuvant therapy against ER-positive breast cancer, and acts as a selective estrogen receptor modulator (SERM). Tamoxifen exhibits highly complex metabolism mediated by several Phase I and II enzymes, including CYP2D6, CYP2C9, CYP2C19, CYP3A4/5 and UDP-glucuronosyltransferases (UGTs), which results in the formation of multiple primary and secondary metabolites that exhibit variable potencies towards estrogen receptors.[1–3] Its therapeutic activity is mainly due to its principal active metabolites 4-hydroxytamoxifen (4-OH-tamoxifen) and endoxifen, which exhibit equivalent ERα binding affinities[4].

Despite the generally successful outcomes of tamoxifen treatment in an adjuvant setting, about 30% of patients show suboptimal clinical response to tamoxifen and eventually succumb to the malignancy. Maintenance of endoxifen levels above certain therapeutic thresholds has recently been suggested to be of critical importance in achieving successful outcomes to tamoxifen[5–8]. However, steady-state endoxifen concentrations in tamoxifen-treated patients are characterized by extremely high interpatient variability and are suspected to be under strong genetic control.

Given the key role of CYP2D6 in mediating the formation of endoxifen, polymorphic *CYP2D6* alleles responsible for amino acid alterations of the CYP2D6 enzyme function (e.g. *CYP2D6**4 [splice acceptor polymorphism], *10 [resulting in p. Pro34Ser substitution] and *41 [defined by rs28371725 causing reduced activity]) have been postulated to predict steady state endoxifen levels and clinical response to tamoxifen. The current scientific state-of-the-art established from various study populations[5,6,9] and modeling analysis[10] as well as thorough discussion of the available literature[11–14] highlighted *CYP2D6* alleles to account for 30-40% of the inter-patient variability in endoxifen concentrations. Yet, other potentially relevant factors remain to be determined with few genome-wide association studies done to date[15,16]. In light of the known inter-ethnic variation of the frequency of functional *CYP2D6* alleles such as CYP*2D6*4*, *CYP2D6*10*, and *CYP2D6*41,* cross-ancestry analysis is in demand to further advance the field. To this end, we previously reported the first cross-ancestry GWAS analysis on the topic and identified the extended CYP2D6 locus as the principle genetic determinant of endoxifen plasma concentrations.[17]

Here we present a largely expanded unbiased cross-ancestry GWAS of more than 2200 breast cancer patients of European (Austria, France, Germany, United Kingdom (UK)), Middle Eastern (Lebanon), and Asian (Singapore) descent to pursue and validate genetic determinants of tamoxifen pharmacokinetics and their potential contribution to breast cancer outcomes (disease free survival, relapse free survival and distant relapse free survival) following tamoxifen treatment.

ARTICLE IN PRESS

ARTICLE IN PRESS

Results

GWAS identified TCF20 rs932376 as additional genetic marker of tamoxifen pharmacokinetics

For the discovery GWAS analysis, patients from Singapore, Germany and Lebanon were clearly grouped according to their self-reported ancestries on principal component analysis of their genome-wide genotyping data (Asians for Singapore, Europeans for Germany, and Middle Eastern for Lebanon, **Figure S2A**). We observed that persons carrying *CYP2D6* poor metabolizer alleles were significantly enriched in European ancestry patients (Germany, 15 out of 286, 5.2%) and Middle Eastern ancestry patients (Lebanon, 3 out of 70, 4.3%) compared to Asian ancestry patients (Singapore, none out of 280, 0%) (**Table 1**).

To perform a genome-wide search for significant genetic determinants of tamoxifen pharmacokinetics, we measured the association between each SNP genotype and log-transformed serum 4-OH-tamoxifen and endoxifen levels at steady-state using linear regression. We additionally adjusted for the top principal components of genetic ancestry to account for ancestry differences between samples. We meta-analyzed the genome-wide association results from all three collections to look for generalizable, robustly significant genetic determinants of serum 4-OH-tamoxifen and endoxifen levels. The meta-analyses showed minimal inflation of test statistics, suggesting the genome-wide association analysis was unlikely to be confounded by unaccounted systematic biases such as population stratification (**Figure S2B**).

We did not observe any genetic marker showing association surpassing genome-wide significance for 4-OH-tamoxifen levels on this discovery GWAS meta-analysis (**Figure S2C**), suggesting that no major genetic locus determines serum 4-OH-tamoxifen levels at steady-state. Turning to the analysis of serum endoxifen levels at steady-state, we note genome-wide significant evidence of association between a cluster of SNP markers at a broad locus mapping to chromosome 22 (**Figure 1** and **Figure S2C**). This broad locus encompassed, amongst other genes, *CYP2D6* which is known to be strongly associated with Z-endoxifen concentration.

Single variant analysis showed the strongest evidence of association to be at rs932376 A>G. This variant was located approximately 3,000 bp upstream of transcription factor 20 (*TCF20*) and more than 70,000 bp upstream of *CYP2D6* (**Figure 1**). This marker rs932376 A>G was present at high frequencies across all three populations studied, and the association was consistent across all 3 collections with no heterogeneity. This pattern of association is unlike that observed at the well-described *CYP2D6*4*, **10*, and **41* alleles which show substantial inter-cohort heterogeneity of above 66.7%.

TCF20 rs932376 and CYP2D6 metabolizer status were both independent predictors of endoxifen levels.

Univariate analysis suggested that each copy of the *TCF20* rs932376-G allele and *CYP2D6* metabolizer status were correlated with steady state endoxifen levels (**Figures S3-S4** and **Table S3**). In the discovery GWAS, median endoxifen was highest among patients with the *TCF20* rs932376-AA genotype (40.0 nM), followed by patients with the heterozygous AG genotype (29.6 nM) and homozygous GG genotype

ARTICLE IN PRESS

(19.6 nM) (*P*<0.001). Median endoxifen of the *CYP2D6* normal/ultra-rapid, intermediate and poor metabolizer patients were 40.9 nM, 24.3 nM and 9.0 nM, respectively (*P*<0.001). On multivariable analysis with adjustments for age at diagnosis, ethnicity, weight/body mass index and menopausal status, *TCF20* rs932376 A>G and *CYP2D6* metabolizer status were both independent predictors of endoxifen levels (**Figure 2**). In the multivariable model, each additional copy of the *TCF20* rs932376-G allele was associated with a 7.48nM reduction (95% CI -10.58 to -4.39; P=2.16×10^{-6}) in mean endoxifen levels. Compared with *CYP2D6* normal/ultra-rapid metabolizers, the mean endoxifen levels of *CYP2D6* intermediate and poor metabolizers were lower by 13.09 nM (95% CI -17.21 to -8.97; P=4.81×10^{-10}) and 25.69 nM (95% CI -28.31 to -23.08; P=1.33×10^{-82}), respectively (**Table S4**). These findings were similarly observed in the validation collection, with each additional copy of the *TCF20* rs932376-G allele associated with a 6.30nM reduction (95% CI -9.51 to -3.09; *P*=0.0001) in mean endoxifen levels. Compared with *CYP2D6* normal/ultra-rapid metabolizers, the mean endoxifen levels of *CYP2D6* intermediate and poor metabolizers were lower by 12.92 nM (95% CI -16.33 to -9.52; P=1.05×10^{-13}) and 26.65 nM (95% CI -29.56 to -23.73; P=9.79×10^{-72}), respectively (**Table S4**).

CYP2D6 metabolizer status was a stronger and more accurate predictor of endoxifen levels than TCF20 rs932376

As *TCF20* rs932376 A>G and *CYP2D6* metabolizer status both showed evidence of association with serum endoxifen levels independent of one another, we compared their predictive value on endoxifen levels. Based on the discovery GWAS, *CYP2D6* metabolizer status accounted for a higher percentage of the predictive information contained in a bivariate model including *TCF20* rs932376 A>G and *CYP2D6* metabolizer status as compared with *TCF20* rs932376 A>G (91.2% vs 48.8%, **Figure 3A**). When we compared mean squared errors (MSE) between prediction models for endoxifen levels, the addition of *CYP2D6* metabolizer status to the model with only *TCF20* rs932376 A>G significantly improved the prediction accuracy of the resulting bivariate model with MSE reduced from 497.5 to 451 (*P*<0.0001). In contrast, the addition of *TCF20* rs932376 A>G to the model with only *CYP2D6* metabolizer status did not significantly improve the prediction accuracy of the resulting bivariate model (MSE reduced marginally from 458.2 to 451, *P*=0.0977, **Figure 3B**). All these findings were similarly observed in the validation collection, and they supported the notion that *CYP2D6* metabolizer status was a stronger and more accurate predictor of endoxifen levels than *TCF20* rs932376 A>G.

TCF20 rs932376 and CYP2D6 metabolizer status were not independent prognostic factors of breast cancer outcomes

Finally, we evaluated whether *TCF20* rs932376 A>G and *CYP2D6* metabolizer status were associated with outcomes in 1326 breast cancer patients who have undergone tamoxifen therapy on an adjuvant setting (**Table 2**). On univariate analysis, *CYP2D6* normal/ultra-rapid and intermediate metabolizers had better DFS (*P*=0.027) and RFS (*P*=0.019) than the poor metabolizers. There were no distinct differences in each endpoint across the *TCF20* rs932376 wildtype (AA), heterozygous (AG), and homozygous (GG) patients (**Figures S5-S7**). On multivariable analysis adjusting for age at diagnosis, ethnicity, body mass index, tumour size, nodal status and tumour grade, neither *TCF20* rs932376 per copy of the G allele nor *CYP2D6* metabolizer status were significantly associated with RFS, DFS and DRFS, albeit a trend for

ARTICLE IN PRESS

decreased RFS was observed for *CYP2D6* poor metabolizers (*P*=0.088) (**Figure 4** and **Tables S5-S7**).

Sensitivity analyses

We conducted two sensitivity analyses. The first investigated the impact of information loss resulting from the categorisation of *CYP2D6* metabolizer status in the multivariable analyses of endoxifen levels and breast cancer outcomes. We reanalysed **Figure 2** and **Figure 4** using *CYP2D6* activity score (**Figure S8 and S9**), and these analyses did not fundamentally change the conclusions of our main analyses. The second sensitivity analysis assessed whether the associations of *TCF20* rs932376 A>G and *CYP2D6* metabolizer status with each breast cancer outcome were dependent on menopausal status. We reanalysed **Figure 4** using subgroup interaction analysis (**Figure S10**). There were no significant differences in the associations of *TCF20* rs932376 A>G and *CYP2D6* metabolizer status with all breast cancer outcomes between the premenopausal and postmenopausal women (*P* of all interaction terms > 0.05).

ARTICLE IN PRESS

ARTICLE IN PRESS

Discussion

Our study is the first multi-ancestry GWAS attempting to survey the genetic determinants underlying tamoxifen pharmacokinetics and clinical outcome. In this extended multi-ancestry GWAS we reconfirmed the genome-wide significant association between a broad locus on chromosome 22 and serum endoxifen concentration[15,17]. This locus contained at least two independent genetic determinants of serum endoxifen levels; the *CYP2D6* gene and *TCF20* rs932376 A>G. Despite a substantial correlation (~39%) between *TCF20* rs932376 A>G and *CYP2D6* metabolizer status, *TCF20* rs932376 A>G remained significantly associated with serum endoxifen levels after adjustment for *CYP2D6* metabolizer status in both the discovery GWAS and validation sample collections. These latter collections were independently ascertained from five countries (France, Germany, Lebanon, Singapore, and the UK) with diverse patient ancestries, but consistency of the association across each constituent site and full adjustment of potentially confounding model covariates suggest validity of the identified *TCF20* association (**Table S4**). Our unbiased multi-ancestry GWAS did not detect any evidence of association surpassing $P < 1\text{x}10^{-6}$ at other candidate loci outside *CYP2D6* and its surrounding region suggesting that common genetic variation at other loci may have minimal contributions to serum endoxifen levels.

Previous studies of the genomic influence on endoxifen concentrations mainly focused on *CYP2D6* due to its functional role in tamoxifen metabolism. However ethnic-specific allele frequency differences and other genetic and non-genetic determinants may have confounded the strength of the association. [27] To this end, it is important to identify markers that may contribute to the prediction of an unexplained 50 to 70% variability of endoxifen plasma levels. The *TCF20* rs932376 A>G variant shows substantial correlation to CYP2D6 activity, and *TCF20* encodes a transcriptional co-regulator which is ubiquitously expressed and has been shown to either enhance or repress the transcriptional activities of other transcription factors, including Sp1 (specificity protein 1), c-Jun, Ets-1 (E twenty-six 1), AR (androgen receptor), SNURF (small nuclear RING-finger)/RNF4, ERα (estrogen receptor α) and Pax6 (paired box protein 6).[28–32] Although these transcription factors are structurally and functionally distinct and bind to different target sequences in the promoter and enhancer regions of the regulated genes, none of them have been reported to specifically regulate genes involved in tamoxifen metabolism including *CYP2D6.*[33] Moreover, given the lack of evidence that *TCF20* rs932376 A>G and its three linked SNPs (rs5751245, rs5758666, rs5751247) directly regulate *TCF20* expression, a cis-acting effect of rs932376 on *CYP2D6* expression is more plausible to explain the strong association with endoxifen pharmacokinetics. This is in line with previous data that identified rs5751247, i.e. one of the three SNPs linked with *TCF20* rs932376 (r^2 =0.93) to be associated with lower CYP2D6 activity and expression.[34] Given its high frequency across ethnicities, we speculate that the *TCF20* rs932376 polymorphism (or its linked SNPs) may have evolved as a *CYP2D6*-controling locus prior to the origin of ethnic-specific *CYP2D6* variants. Its role in modulating serum endoxifen levels, and how they interfere possibly via long-distance enhancers/repressors on *CYP2D6* will require further mechanistic investigations. Nevertheless, its functional characterization with respect to its role in modulating serum endoxifen levels remains unclear and needs to be further elucidated.

ARTICLE IN PRESS

Despite their significant effect on modulating serum endoxifen levels, both *TCF20* rs932376 A>G and CYP2D6 metabolizer status were not significantly associated with breast cancer outcomes in a multivariable model adjusting for age, ethnicity, body mass index, tumour size, nodal status, and tumour grade. However, we observed a trend of shorter DFS and RFS (univariate) and worse RFS (multivariate) for CYP2D6 poor metabolizer patients compared to those with normal metabolism, in line with a possible modulatory impact of CYP2D6 metabolizer status. To this end it is important to note that a recent systematic review and meta-analysis of 33 retrospective cohorts suggested that *CYP2D6* metabolizer status had a modest effect on recurrence risk. Upon bias correction for loss of heterozygosity and incomplete genotyping, the findings highlighted the limited utility of retrospective studies for broad pharmacogenomic analyses.[35] Of note, novel approaches to overcome impaired tamoxifen metabolism using either single-agent endoxifen[36] or endoxifen supplementation[37] have been reported.

This study is not without limitations. The GWAS approach includes only common tagging variants, most of which are non-coding and may have unknown functional consequences and thus may not be sufficiently informative or tractable, including that of *TCF20* rs932376. Rare genetic variants with minor allele frequencies of less than one percent may have escaped capture using this approach. Nevertheless, the genetic effects observed for this common *TCF20* rs932376 marker on stratification of serum endoxifen levels were both highly significant and validated across multiple populations. The challenge remains however, to provide a mechanism linking this variant to inter-individual endoxifen variability. Another limitation is that tamoxifen adherence and concomitant medication use were not included as covariates in our analyses due to incomplete information in some of the cohorts studied which may have impacted the results. This study was also limited by the under-representation of African ancestry/Hispanics ethnic groups, who were not present in large numbers in the cohorts studied. We were also unable to perform any direct association analyses between serum endoxifen concentrations and BC survival outcomes due to the unavailability of endoxifen data in two out of four survival cohorts evaluated.

In conclusion, this study identified *TCF20* rs932376 A>G as an additional genetic factor that significantly associates with steady-state serum endoxifen levels. Despite its minor contribution when compared to the *CYP2D6* that convincingly and plausibly associates with endoxifen levels, our findings clearly advance the current understanding of the pharmacogenetic architecture underlying the variability in tamoxifen pharmacokinetics. However, this study has shown insufficient evidence to support routine assessment of pharmacogenetic factors to predict pharmacodynamic outcome measures. The findings of this study also highlight the need of large uniformly collected patient cohorts that follow strict inclusion criteria, avoid data bias and include drug metabolite levels to further elucidate the role of pharmacokinetic variability on tamoxifen response.

ARTICLE IN PRESS

ARTICLE IN PRESS

Methods

Breast cancer patient collections

Hormone-receptor positive breast cancer patients who took 20mg tamoxifen daily for at least 8 weeks prior to pharmacokinetics blood sampling were included in the pharmacokinetic GWAS study (**Figure S1A and Table S1**). The discovery GWAS comprised patients from Singapore, Lebanon, and Germany. Validation of significant associations arising from the discovery GWAS was sought in three independently ascertained cohorts from the UK, France and Singapore. Hormone-receptor positive breast cancer patients adjuvantly treated with tamoxifen were ascertained from Singapore, UK, Germany and Austria for the analysis of breast cancer outcomes (**Figure S1B** and **Table S2**). Study inclusion and exclusion criteria for all cohorts included in this study have been described previously.[11,18–21] The study was carried out in accordance with the provisions of the Declaration of Helsinki of 1975. Institutional ethics approval were obtained from the Singapore Health Services' Centralised Institutional Review Board (Ref: 2008/402/B and 2020/2156/B) (Singapore), the American University of Beirut institutional ethics review board (Lebanon), the Ethics Commission of the University of Tuebingen (Ref: 340/2004) and respective local ethics committees of all participating centres in Germany and in France, the Ethical Committee of the Medical University Graz (Austria), and South and West MultiCentre Research Ethics committee (MREC 00/6/69) (UK).

Measurements of steady-state serum levels of tamoxifen and the active metabolites 4OH tamoxifen and endoxifen

Blood samples (3 ml) were drawn from tamoxifen-treated patients after at least 8 weeks of tamoxifen therapy (20 mg per day). Plasma extraction was performed by centrifugation under light protection within 30 minutes of venipuncture and stored at -80°C until analysis. Steady-state plasma concentrations of tamoxifen and its active metabolites, 4-hydroxytamoxifen and endoxifen in their active (Z)-isomers, were quantified as previously described.[5,22]

Discovery genome-wide genotyping for genome-wide association study (GWAS)

Genome-wide genotyping for the discovery collection (comprising patients from Singapore (N=280), Germany (N=286) and Lebanon (N=70)) (**Table S1**) with serum tamoxifen and active metabolite levels collected at steady state was performed using the Illumina OmniExpress beadchips, following manufacturer's instructions (www.illumina.com).

Discovery GWAS data quality control and analysis to identify key genetic marker of tamoxifen pharmacokinetics

A total of 595,589 autosomal SNP markers passed stringent quality control checks (completeness in genotyping exceeding 95%, minor allele frequency exceeding 1 percent, with no significant deviation from Hardy-Weinberg equilibrium [*P*-value for deviation > $1x10^{-6}$]) in all three collections. These SNP markers were brought forward for further association analysis. We subjected each individual sample to similar quality checks for completeness of genotyping, extremes of heterozygosity indicative of sample cross-contamination, as well as examination for possible relatedness amongst the samples genotyped.

ARTICLE IN PRESS

Analysis of their genetic ancestry using principal component analysis showed minimal stratification within the individual collections (Asian, European ancestry, and Middle Eastern). Association between the active tamoxifen metabolites (4-OH-tamoxifen and endoxifen) were measured using linear regression, with the serum metabolite levels analysed as a quantitative trait with log-transformation. We included the top 5 principal components of genetic stratification into the linear regression model to compensate for residual population stratification. The widely described $P < 5 \times 10^{-8}$ was pre-set as the threshold for genome-wide significance.

Validation of genome-wide genotyping

For the validation collection, patients from UK-POSH (N=262), France-PHACS (N=535) and Singapore-NCC2003 (N=72) studies (**Table S1**) were subjected to genome-wide genotyping based on the Illumina global screening array. The identified key genetic marker from discovery GWAS, *TCF20* rs932376 A>G, was extracted from the genome-wide genotyping data for all three cohorts and analyzed in association with endoxifen levels at steady state.

Statistical analysis

CYP2D6 metabolizer status was determined using the activity score (AS) based on the final consensus CYP2D6 genotype to phenotype translation guidelines published in 2020 by the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenomics Working Group[23]: patients with poor metabolizer status had an activity score of 0, patients with intermediate metabolizer status had an activity score of between 0.25 and 1, whereas patients with normal metabolizer status had an activity score of 1.25 – 2.25. Patients with ultra-rapid metabolizer status had an activity score >2.25.[23] Comparison of the distribution of endoxifen levels by *CYP2D6* metabolizer status and by *TCF20* rs932376 A>G genotype status were compared using either Wilcoxon rank sum test (2 groups) or Kruskal Wallis test (>2 groups).

The association between endoxifen levels and *CYP2D6* metabolizer status, *TCF20* rs932376 A>G per-copy of the minor G allele, as well as the covariates age at diagnosis, ethnicity, weight/body mass index and menopausal status were evaluated using Gaussian generalized linear model with log link function. Average marginal effects were derived to express the effect of each covariate in the fitted multivariable model on the adjusted change in mean endoxifen levels. Diagnostic checks of fitted models were performed by assessing normality of residuals. Likelihood ratio testing of nested models were performed to compare the predictive value of *CYP2D6* metabolizer status and *TCF20* rs932376 A>G per-copy of the G allele on endoxifen levels. An adequacy index (AI) based on log-likelihood ratio statistics was used to quantify the proportion of predictive information contained in a model with both these predictors that is attributable to each predictor[24]. Bias-corrected confidence intervals for AI were estimated based on bootstrapping with 1000 resamples. Differences in the average of squared residuals between pairs of prediction models for endoxifen levels were tested using one sample T-tests.

We examined disease free survival (DFS), relapse free survival (RFS) and distant relapse free survival (DRFS) as breast cancer outcomes, with each of these endpoints defined according to STEEP 2.0.[25] DFS was estimated based on Kaplan-Meier method and compared using log-rank test. Cox proportional hazards regression models were used to assess the association between DFS and various covariates.

ARTICLE IN PRESS

Covariates with univariate $P<0.1$ were selected for inclusion in the multivariable analysis. To account for competing risks, RFS and DRFS were estimated based on 1 – cumulative incidence function, compared using Gray's test and their association with covariates assessed based on Fine and Gray regression model. The proportional hazards assumption was verified based on Schoenfeld's residuals.

Analyses were carried out using SAS v9.4 (SAS Institute Inc., Cary, NC), Stata v16 (StataCorp, College Station, TX) and R v4.4.1. All statistical tests were two-sided with a 5% significance level. There was minimal missing data in this study; no imputation for missing values was performed. The STREGA reporting guidelines were used.[26]

ARTICLE IN PRESS

ARTICLE IN PRESS

Data availability

Data supporting this study have been de-identified and only known to the respective research teams involved in patient recruitment. The data is not publicly available due to confidentiality of the research participants and is governed by the data sharing policies of the respective research institutions. Please contact the corresponding author (BC) for any request.

Acknowledgements

This study was supported by grants from the National Medical Research Council Singapore (NMRC/1159/2008, NMRCB1011, NRFCG1516, NMRCG13163 and NMRC/CIRG/1423/2015, MOH-000377), the National Research Foundation of Singapore grant (NRF-NRFI2018-01); the Robert Bosch Foundation, Stuttgart, Germany and German Federal Ministry of Education and Research (BMBF; 01ZP0502, 01EK1509A); the Deutsche Foschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2180–390900677 and DFG grants SCHR 1323/2-1 and MU 1727/2-1; the German Cancer Consortium (DKTK), Partner Site Tübingen, 72074 Tübingen and the Interfaculty Center for Pharmacogenomics and Drug Research (ICEPHA), University of Tuebingen (no grant number). The funding sources had no involvement in the study design, collection, analysis and interpretation of data, writing of the manuscript and decision for publication. Additional funding for this study was provided by the Biomedical Research Council, Agency for Science, Technology and Research, Singapore (no grant number). The funders had no role in the study design, collection, analysis and interpretation of data as well as drafting and submission of this manuscript.

Authors' contributions

BC conceived and designed the study. EHL, EC, SC, TEM, JEA, NS, JSLL, SSL, CA, ZL, KSS, DME, WJT, TM, NKZ, SW, BG, LD, QG, CC, HME, LH, JD, RCHN, YSY, MW, FYW, NSW, PACS, RD, PK, UL, TL, AT, ES, ME, UH, PAF, MWB, FD, MWK, HBB, WS, WR, MS were involved in patient collection, conduct of study and data collection. CCK, WSO, TEM, HBB, MS, FT and BC contributed to data analysis and interpretation. CCK, WSO and BC drafted the manuscript. All authors participated in manuscript editing, review and approval of final revised manuscript.

Competing interests

Carlos Caldas is a member of the AstraZeneca External Science Panel and has received research grants (administered by the University of Cambridge) from AstraZeneca, Genentech, Roche and Servier. Helena Earl has received research grants from Roche and Sanofi-Aventis (administered by Cambridge University Hospital Trust), honoraria and travel expenses from Daiichi-Sankyo, Astra Zeneca, Pfizer, Amgen and Prime Oncology all outside the submitted work. Jean Abraham has received research grants from AstraZeneca (administered by the Cambridge University Hospital Trust and the University of Cambridge, Dept of Oncology), honoraria and travel expenses from Astra Zeneca and Pfizer all outside the submitted work. All other authors do not have a competing interest.

ARTICLE IN PRESS

ARTICLE IN PRESS

References

1. Lim, J. S. L. *et al.* Impact of CYP2D6, CYP3A5, CYP2C9 and CYP2C19 polymorphisms on tamoxifen pharmacokinetics in Asian breast cancer patients. *Br. J. Clin. Pharmacol.* **71**, 737–750 (2011).
2. Lim, J. S. L. *et al.* Association of CYP2C19*2 and associated haplotypes with lower norendoxifen concentrations in tamoxifen-treated Asian breast cancer patients. *Br. J. Clin. Pharmacol.* **81**, 1142–1152 (2016).
3. Sutiman, N. *et al.* Pharmacogenetics of UGT1A4, UGT2B7 and UGT2B15 and Their Influence on Tamoxifen Disposition in Asian Breast Cancer Patients. *Clin. Pharmacokinet.* **55,** 1239–1250 (2016).
4. Lim, Y. C., Desta, Z., Flockhart, D. A. & Skaar, T. C. Endoxifen (4-hydroxy-N-desmethyl-tamoxifen) has anti-estrogenic effects in breast cancer cells with potency similar to 4-hydroxy-tamoxifen. *Cancer Chemother. Pharmacol.* **55**, 471–478 (2005).
5. Mürdter, T. E. *et al.* Activity levels of tamoxifen metabolites at the estrogen receptor and the impact of genetic polymorphisms of phase I and II enzymes on their concentration levels in plasma. *Clin. Pharmacol. Ther.* **89**, 708–717 (2011).
6. Madlensky, L. *et al.* Tamoxifen metabolite concentrations, CYP2D6 genotype, and breast cancer outcomes. *Clin. Pharmacol. Ther.* **89**, 718–725 (2011).
7. Gong, I. Y. *et al.* Determination of clinically therapeutic endoxifen concentrations based on efficacy from human MCF7 breast cancer xenografts. *Breast Cancer Res. Treat.* **139**, 61–69 (2013).
8. Teft, W. A. *et al.* CYP3A4 and seasonal variation in vitamin D status in addition to CYP2D6 contribute to therapeutic endoxifen level during tamoxifen therapy. *Breast Cancer Res. Treat.* **139**, 95–105 (2013).
9. Saladores, P. *et al.* Tamoxifen metabolism predicts drug concentrations and outcome in premenopausal patients with early breast cancer. *Pharmacogenomics J.* **15**, 84–94 (2015).
10. Schroth, W. *et al.* Improved Prediction of Endoxifen Metabolism by CYP2D6 Genotype in Breast Cancer Patients Treated with Tamoxifen. *Front. Pharmacol.* **8**, 582 (2017).
11. Lim, J. S. L. *et al.* Impact of CYP2D6, CYP3A5, CYP2C9 and CYP2C19 polymorphisms on tamoxifen pharmacokinetics in Asian breast cancer patients. *Br. J. Clin. Pharmacol.* **71**, 737–750 (2011).
12. Goetz, M. P. & Ingle, J. N. CYP2D6 genotype and tamoxifen: considerations for proper nonprospective studies. *Clin. Pharmacol. Ther.* **96**, 141–144 (2014).
13. Rae, J. M., Regan, M., Leyland-Jones, B., Hayes, D. F. & Dowsett, M. CYP2D6 genotype should not be used for deciding about tamoxifen therapy in postmenopausal breast cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **31**, 2753–2755 (2013).
14. Brauch, H. & Schwab, M. Prediction of tamoxifen outcome by genetic variation of CYP2D6 in post-menopausal women with early breast cancer. *Br. J. Clin. Pharmacol.* **77**, 695–703 (2014).
15. Sanchez-Spitman, A. B. *et al.* A Genome-Wide Association Study of Endoxifen Serum Concentrations and Adjuvant Tamoxifen Efficacy in Early-Stage Breast Cancer Patients. *Clin. Pharmacol. Ther.* **116**, 155–164 (2024).
16. Hennig, E. E. *et al.* Non-CYP2D6 Variants Selected by a GWAS Improve the Prediction of Impaired Tamoxifen Metabolism in Patients with Breast Cancer. *J. Clin. Med.* **8**, 1087 (2019).

ARTICLE IN PRESS

ARTICLE IN PRESS

17. Khor, C. C. *et al.* Cross-Ancestry Genome-Wide Association Study Defines the Extended CYP2D6 Locus as the Principal Genetic Determinant of Endoxifen Plasma Concentrations. *Clin. Pharmacol. Ther.* **113**, 712–723 (2023).
18. Mürdter, T. E. *et al.* Activity levels of tamoxifen metabolites at the estrogen receptor and the impact of genetic polymorphisms of phase I and II enzymes on their concentration levels in plasma. *Clin. Pharmacol. Ther.* **89**, 708–717 (2011).
19. Awada, Z. *et al.* Pharmacogenomics variation in drug metabolizing enzymes and transporters in relation to docetaxel toxicity in Lebanese breast cancer patients: paving the way for OMICs in low and middle income countries. *Omics J. Integr. Biol.* **17**, 353–367 (2013).
20. Eccles, D. *et al.* Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH): study protocol. *BMC Cancer* **7**, 160 (2007).
21. Puszkiel, A. *et al.* Factors Affecting Tamoxifen Metabolism in Patients With Breast Cancer: Preliminary Results of the French PHACS Study. *Clin. Pharmacol. Ther.* **106**, 585–595 (2019).
22. Arellano, C., Allal, B., Goubaa, A., Roché, H. & Chatelut, E. An UPLC-MS/MS method for separation and accurate quantification of tamoxifen and its metabolites isomers. *J. Pharm. Biomed. Anal.* **100**, 254–261 (2014).
23. Caudle, K. E. *et al.* Standardizing CYP2D6 Genotype to Phenotype Translation: Consensus Recommendations from the Clinical Pharmacogenetics Implementation Consortium and Dutch Pharmacogenetics Working Group. *Clin. Transl. Sci.* **13**, 116–124 (2020).
24. Harrell, F. E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* (Springer, New York, 2001).
25. Tolaney, S. M. *et al.* Updated Standardized Definitions for Efficacy End Points (STEEP) in Adjuvant Breast Cancer Clinical Trials: STEEP Version 2.0. *J. Clin. Oncol.* **39**, 2720–2731 (2021).
26. Little, J. *et al.* STrengthening the REporting of Genetic Association Studies (STREGA)--an extension of the STROBE statement. *Genet. Epidemiol.* **33**, 581–598 (2009).
27. Binkhorst, L., Mathijssen, R. H. J., Jager, A. & van Gelder, T. Individualization of tamoxifen therapy: much more than just CYP2D6 genotyping. *Cancer Treat. Rev.* **41**, 289–299 (2015).
28. Rekdal, C., Sjøttem, E. & Johansen, T. The nuclear factor SPBP contains different functional domains and stimulates the activity of various transcriptional activators. *J. Biol. Chem.* **275**, 40288–40300 (2000).
29. Sanz, L., Moscat, J. & Diaz-Meco, M. T. Molecular characterization of a novel transcription factor that controls stromelysin expression. *Mol. Cell. Biol.* **15**, 3164–3170 (1995).
30. Lyngsø, C. *et al.* Interaction between the transcription factor SPBP and the positive cofactor RNF4. An interplay between protein binding zinc fingers. *J. Biol. Chem.* **275**, 26144–26149 (2000).
31. Elvenes, J. *et al.* Pax6 represses androgen receptor-mediated transactivation by inhibiting recruitment of the coactivator SPBP. *PloS One* **6**, e24659 (2011).
32. Gburcik, V., Bot, N., Maggiolini, M. & Picard, D. SPBP is a phosphoserine-specific repressor of estrogen receptor alpha. *Mol. Cell. Biol.* **25**, 3421–3430 (2005).
33. Pan, X., Ning, M. & Jeong, H. Transcriptional Regulation of CYP2D6 Expression. *Drug Metab. Dispos. Biol. Fate Chem.* **45**, 42–48 (2017).
34. Yang, X. *et al.* Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* **20**, 1020–1036 (2010).

ARTICLE IN PRESS

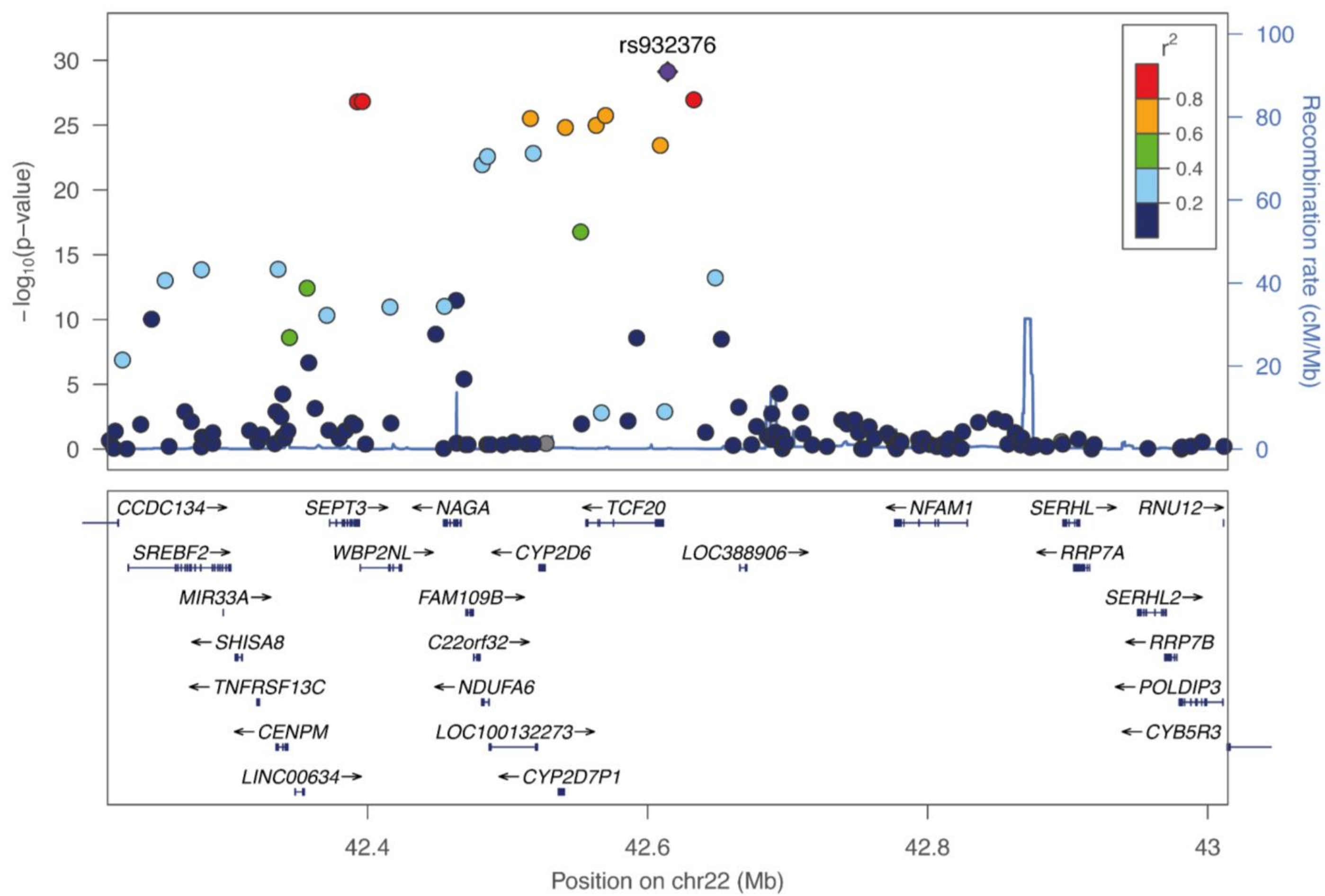
ARTICLE IN PRESS

35. MacLehose, R. F., Ahern, T. P., Collin, L. J., Li, A. & Lash, T. L. *CYP2D6* Phenotype and Breast Cancer Outcomes: A Bias Analysis and Meta-Analysis. *Cancer Epidemiol. Biomarkers Prev.* **34**, 224–233 (2025).
36. Goetz, M. P. *et al.* First-in-Human Phase I Study of the Tamoxifen Metabolite Z-Endoxifen in Women With Endocrine-Refractory Metastatic Breast Cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **35**, 3391–3400 (2017).
37. Mürdter, T. E. *et al.* Supplementation of Tamoxifen with low-dose Endoxifen in breast cancer patients with impaired tamoxifen metabolism (TAMENDOX): a randomized controlled phase 1/2 trial. *Clin. Cancer Res.* **31**, 4903–4911 (2025)

ARTICLE IN PRESS

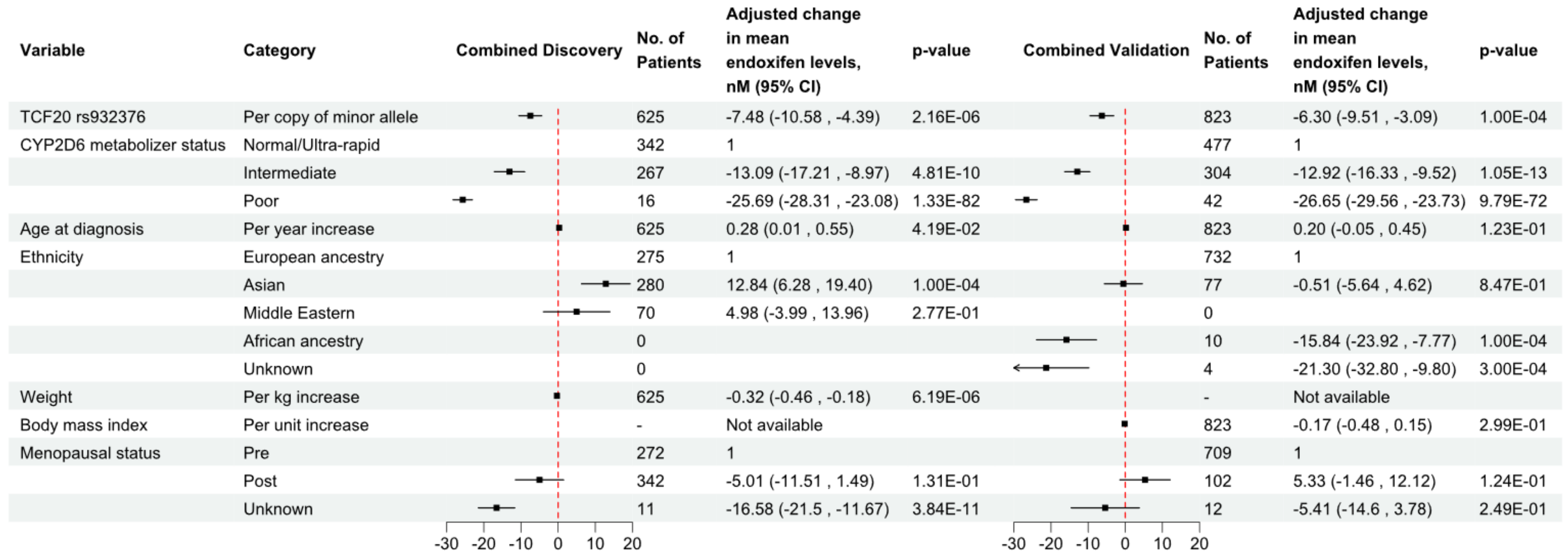
ARTICLE IN PRESS

Figure 1



ARTICLE IN PRESS

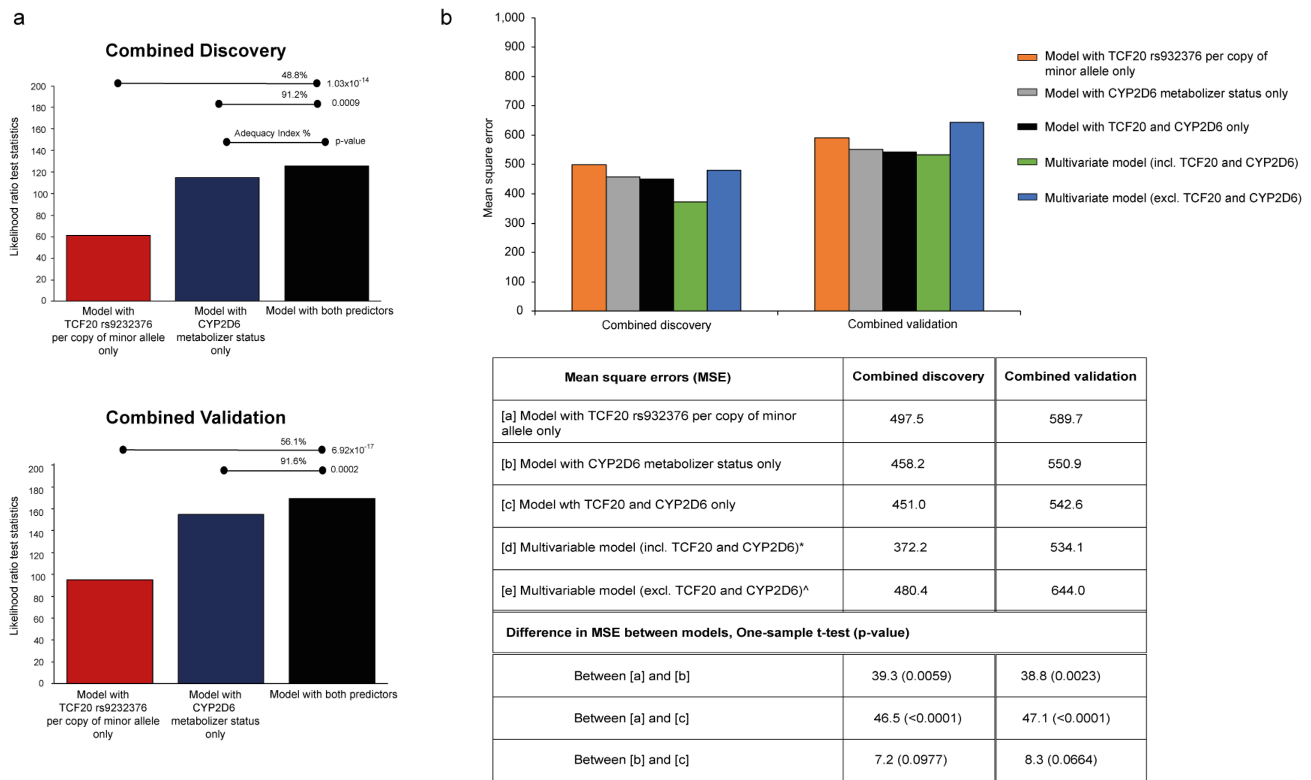
Figure 2



Variable	Category	Combined Discovery	No. of Patients	Adjusted change in mean endoxifen levels, nM (95% CI)	p-value	Combined Validation	No. of Patients	Adjusted change in mean endoxifen levels, nM (95% CI)	p-value
TCF20 rs932376	Per copy of minor allele		625	-7.48 (-10.58 , -4.39)	2.16E-06		823	-6.30 (-9.51 , -3.09)	1.00E-04
CYP2D6 metabolizer status	Normal/Ultra-rapid		342	1			477	1	
	Intermediate		267	-13.09 (-17.21 , -8.97)	4.81E-10		304	-12.92 (-16.33 , -9.52)	1.05E-13
	Poor		16	-25.69 (-28.31 , -23.08)	1.33E-82		42	-26.65 (-29.56 , -23.73)	9.79E-72
Age at diagnosis	Per year increase		625	0.28 (0.01 , 0.55)	4.19E-02		823	0.20 (-0.05 , 0.45)	1.23E-01
Ethnicity	European ancestry		275	1			732	1	
	Asian		280	12.84 (6.28 , 19.40)	1.00E-04		77	-0.51 (-5.64 , 4.62)	8.47E-01
	Middle Eastern		70	4.98 (-3.99 , 13.96)	2.77E-01		0		
	African ancestry		0				10	-15.84 (-23.92 , -7.77)	1.00E-04
	Unknown		0				4	-21.30 (-32.80 , -9.80)	3.00E-04
Weight	Per kg increase		625	-0.32 (-0.46 , -0.18)	6.19E-06		-	Not available	
Body mass index	Per unit increase		-	Not available			823	-0.17 (-0.48 , 0.15)	2.99E-01
Menopausal status	Pre		272	1			709	1	
	Post		342	-5.01 (-11.51 , 1.49)	1.31E-01		102	5.33 (-1.46 , 12.12)	1.24E-01
	Unknown		11	-16.58 (-21.5 , -11.67)	3.84E-11		12	-5.41 (-14.6 , 3.78)	2.49E-01

ARTICLE IN PRESS

Figure 3



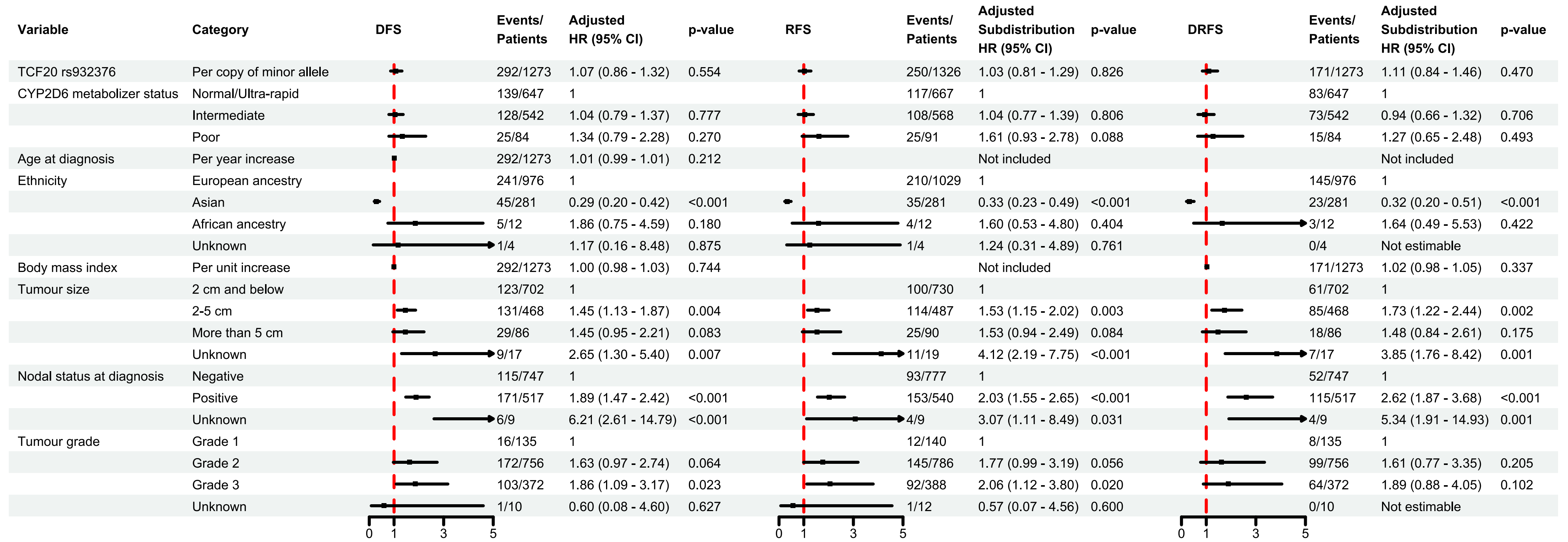
Mean square errors (MSE)	Combined discovery	Combined validation
[a] Model with TCF20 rs932376 per copy of minor allele only	497.5	589.7
[b] Model with CYP2D6 metabolizer status only	458.2	550.9
[c] Model wth TCF20 and CYP2D6 only	451.0	542.6
[d] Multivariable model (incl. TCF20 and CYP2D6)*	372.2	534.1
[e] Multivariable model (excl. TCF20 and CYP2D6)^	480.4	644.0
Difference in MSE between models, One-sample t-test (p-value)		
Between [a] and [b]	39.3 (0.0059)	38.8 (0.0023)
Between [a] and [c]	46.5 (<0.0001)	47.1 (<0.0001)
Between [b] and [c]	7.2 (0.0977)	8.3 (0.0664)

* Variables included were TCF20, CYP2D6, age, ethnicity, weight/body mass index and menopausal status.
^ Variables included were age, ethnicity, weight/body mass index and menopausal status.

ARTICLE IN PRESS

ARTICLE IN PRESS

Figure 4



Variable	Category	DFS Events/ Patients	Adjusted HR (95% CI)	p-value	RFS Events/ Patients	Adjusted Subdistribution HR (95% CI)	p-value	DRFS Events/ Patients	Adjusted Subdistribution HR (95% CI)	p-value
TCF20 rs932376	Per copy of minor allele	292/1273	1.07 (0.86 - 1.32)	0.554	250/1326	1.03 (0.81 - 1.29)	0.826	171/1273	1.11 (0.84 - 1.46)	0.470
CYP2D6 metabolizer status	Normal/Ultra-rapid	139/647	1		117/667	1		83/647	1	
	Intermediate	128/542	1.04 (0.79 - 1.37)	0.777	108/568	1.04 (0.77 - 1.39)	0.806	73/542	0.94 (0.66 - 1.32)	0.706
	Poor	25/84	1.34 (0.79 - 2.28)	0.270	25/91	1.61 (0.93 - 2.78)	0.088	15/84	1.27 (0.65 - 2.48)	0.493
Age at diagnosis	Per year increase	292/1273	1.01 (0.99 - 1.01)	0.212		Not included			Not included	
Ethnicity	European ancestry	241/976	1		210/1029	1		145/976	1	
	Asian	45/281	0.29 (0.20 - 0.42)	<0.001	35/281	0.33 (0.23 - 0.49)	<0.001	23/281	0.32 (0.20 - 0.51)	<0.001
	African ancestry	5/12	1.86 (0.75 - 4.59)	0.180	4/12	1.60 (0.53 - 4.80)	0.404	3/12	1.64 (0.49 - 5.53)	0.422
	Unknown	1/4	1.17 (0.16 - 8.48)	0.875	1/4	1.24 (0.31 - 4.89)	0.761	0/4	Not estimable	
Body mass index	Per unit increase	292/1273	1.00 (0.98 - 1.03)	0.744		Not included		171/1273	1.02 (0.98 - 1.05)	0.337
Tumour size	2 cm and below	123/702	1		100/730	1		61/702	1	
	2-5 cm	131/468	1.45 (1.13 - 1.87)	0.004	114/487	1.53 (1.15 - 2.02)	0.003	85/468	1.73 (1.22 - 2.44)	0.002
	More than 5 cm	29/86	1.45 (0.95 - 2.21)	0.083	25/90	1.53 (0.94 - 2.49)	0.084	18/86	1.48 (0.84 - 2.61)	0.175
	Unknown	9/17	2.65 (1.30 - 5.40)	0.007	11/19	4.12 (2.19 - 7.75)	<0.001	7/17	3.85 (1.76 - 8.42)	0.001
Nodal status at diagnosis	Negative	115/747	1		93/777	1		52/747	1	
	Positive	171/517	1.89 (1.47 - 2.42)	<0.001	153/540	2.03 (1.55 - 2.65)	<0.001	115/517	2.62 (1.87 - 3.68)	<0.001
	Unknown	6/9	6.21 (2.61 - 14.79)	<0.001	4/9	3.07 (1.11 - 8.49)	0.031	4/9	5.34 (1.91 - 14.93)	0.001
Tumour grade	Grade 1	16/135	1		12/140	1		8/135	1	
	Grade 2	172/756	1.63 (0.97 - 2.74)	0.064	145/786	1.77 (0.99 - 3.19)	0.056	99/756	1.61 (0.77 - 3.35)	0.205
	Grade 3	103/372	1.86 (1.09 - 3.17)	0.023	92/388	2.06 (1.12 - 3.80)	0.020	64/372	1.89 (0.88 - 4.05)	0.102
	Unknown	1/10	0.60 (0.08 - 4.60)	0.627	1/12	0.57 (0.07 - 4.56)	0.600	0/10	Not estimable	

Table 1

	Discovery cohorts			Validation cohorts			Combined discovery	Combined validation
	Singapore-NCC0801	Lebanon	Germany-IKP211	UK-POSH	France-PHACS	Singapore-NCC2003		
No. of patients	280	70	286	262	535	72	636	859
Age at diagnosis, years[1]								
Median (IQR)	47 (44-53)	43 (39-46)	64 (58-69)	37 (35-39)	46 (42 - 50)	44 (40-49)	53.5 (46-64)	42 (38-48)
No. with missing data, n (%)	0	0	0	0	4 (0.7)	0	0	4 (0.5)
Ethnicity, n (%)								
African ancestry	0	0	0	10 (3.8)	0	0	0	10 (1.2)
Asians	280 (100)	0	0	5 (1.9)	0	72 (100)	280 (44)	77 (8.9)
European ancestry	0	0	286 (100)	243 (92.7)	535 (100)	0	286 (45)	778 (89.5)
Middle Eastern	0	70 (100)	0	0	0	0	70 (11)	0
Unknown	0	0	0	4 (1.5)	0	0	0	4 (0.5)
Weight, kg								
Median (IQR)	57.6 (51.0-63.4)	69.0 (61.0-76.0)	71.2 (64.0-79.1)	Not available	63.0 (56.0-72.0)	61.0 (54.1-67.1)	64 (56.3-73.5)	Not tabulated
No. with missing data, n (%)	0	0	11 (3.8)		6 (1.1)	0	11 (1.7)	
Body mass index, kg/m^2								
Median (IQR)	23.2 (21.1-26.2)	25.8 (23.6-29.4)	Not available	24.0 (21.8-27.5)	23.4 (21.0-26.4)	23.2 (21.2-26.8)	Not tabulated	23.5 (21.3-26.8)
Menopausal status, n (%)								
Premenopausal	202 (72.1)	70 (100)	0	262 (100)	459 (85.8)	29 (40.3)	272 (42.8)	750 (86.3)
Postmenopausal	78 (27.9)	0	274 (95.8)	0	67 (12.5)	40 (55.6)	352 (55.3)	107 (12.3)
Unknown	0	0	12 (4.2)	0	9 (1.7)	3 (4.2)	12 (1.9)	12 (1.4)
Overall stage at diagnosis, n (%)								
I	102 (36.4)	23 (32.9)	192 (67.1)	Not available	Not available	27 (37.5)	317 (49.8)	Not tabulated
II	126 (45.0)	32 (45.7)	83 (29.0)			34 (47.2)	241 (37.9)	
III	49 (17.5)	15 (21.4)	9 (3.1)			7 (9.7)	73 (11.5)	
IV	3 (1.1)	0	0			2 (2.8)	3 (0.5)	
Unknown	0	0	2 (0.7)			2 (2.8)	2 (0.3)	
ER status, n (%)								
Positive	257 (91.8)	69 (98.6)	285 (99.7)	Not available	Not available	72 (100)	611 (96.1)	Not tabulated
Negative	23 (8.2)	1 (1.4)	1 (0.3)			0	25 (3.9)	
PR status, n (%)								
Positive	250 (89.3)	56 (80.0)	248 (86.7)	Not available	Not available	66 (91.7)	554 (87.1)	Not tabulated
Negative	30 (10.7)	14 (20.0)	38 (13.3)			6 (8.3)	82 (12.9)	
HER2 status, n (%)								
Positive	68 (24.3)	16 (22.9)	9 (3.1)	52 (19.8)	Not available	0	93 (14.6)	Not tabulated
Negative	206 (73.6)	54 (77.1)	238 (83.2)	117 (44.7)		71 (98.6)	498 (78.3)	
Equivocal	3 (1.1)	0	37 (12.9)	3 (1.1)		0	40 (6.3)	
Unknown	3 (1.1)	0	2 (0.7)	90 (34.4)		1 (1.4)	5 (0.8)	
Tamoxifen setting, n (%)								
Adjuvant	277 (98.9)	70 (100)	286 (100)	262 (100)	535 (100)	70 (97.2)	633 (99.5)	867 (99.8)
Metastatic	3 (1.1)	0	0	0	0	2 (2.8)	3 (0.5)	2 (0.2)
Radiotherapy, n (%)								

ARTICLE IN PRESS

ARTICLE IN PRESS

Yes	183 (65.4)	Not available	Not available	215 (82.1)	Not available	51 (70.8)	Not tabulated	Not tabulated
No	96 (34.3)			47 (17.9)		20 (27.8)		
Unknown	1 (0.4)			0		1 (1.4)		
Chemotherapy, n (%)								
Yes	225 (80.4)	Not available	Not available	195 (74.4)	Not available	39 (54.2)	Not tabulated	Not tabulated
No	55 (19.6)			67 (25.6)		33 (45.8)		
TCF20 rs932376 genotype status, n (%)								
Wild-type	103 (36.8)	36 (51.4)	147 (51.4)	127 (48.5)	272 (50.8)	24 (33.3)	286 (45.0)	423 (48.7)
Heterozygous	134 (47.9)	28 (40.0)	111 (38.8)	108 (41.2)	216 (40.4)	34 (47.2)	273 (42.9)	358 (41.2)
Homozygous	43 (15.4)	6 (8.6)	28 (9.8)	27 (10.3)	47 (8.8)	14 (19.4)	77 (12.1)	88 (10.1)
CYP2D6 metabolizer status, n(%)								
Poor	0	3 (4.3)	15 (5.2)	20 (7.6)	23 (4.3)	1 (1.4)	18 (2.8)	44 (5.1)
Intermediate	130 (46.4)	21 (30.0)	120 (42.0)	109 (41.6)	179 (33.5)	32 (44.4)	271 (42.6)	320 (36.8)
Normal	144 (51.4)	35 (50.0)	143 (50.0)	130 (49.6)	312 (58.3)	39 (54.2)	322 (50.6)	481 (55.4)
Normal to Ultra-rapid	5 (1.8)	0	1 (0.3)	0	0	0	6 (0.9)	0
Ultra-rapid	1 (0.4)	11 (15.7)	7 (2.4)	3 (1.1)	21 (3.9)	0	19 (3.0)	24 (2.8)

IQR, interquartile range; ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2

[1]Age at study inclusion for France-PHACS cohort

Note: Certain variables in the combined cohort were not computed as data was not available in at least one of the cohorts

ARTICLE IN PRESS

Table 2

	Singapore-NCC0801	UK-POSH	Germany-INET	Austria-TIGER	Combined
No. of patients	276	301	525	224	1326
Age at diagnosis, years					
Median (IQR)	48 (44-53)	37 (35-39)	64 (59-71)	61 (51-70)	55 (40-65)
Ethnicity, n (%)					
African ancestry	0	12 (4.0)	0	0	12 (0.9)
Asians	276 (100)	5 (1.7)	0	0	281 (21.2)
European ancestry	0	280 (93.0)	525 (100)	224 (100)	1029 (77.6)
Unknown	0	4 (1.3)	0	0	4 (0.3)
Body mass index, kg/m^2					
Median (range)	23.2 (21.1-26.2)	24.2 (21.8-27.8)	25.8 (23.5-29.0)	25.8 (23.5-30.1)	24.9 (22.5-28.4)
No. with missing data	0	8 (2.7)	9 (1.7)	36 (16.1)	53 (4.0)
Menopausal status, n (%)					
Premenopausal	198 (71.7)	301 (100)	0	52 (23.2)	551 (41.6)
Postmenopausal	78 (28.3)	0	525 (100)	172 (76.8)	775 (58.4)
Tumor size at diagnosis, n (%)					
≤2 cm	146 (52.9)	153 (50.8)	291 (55.4)	140 (62.5)	730 (55.1)
2-5 cm	112 (40.6)	122 (40.5)	202 (38.5)	51 (22.8)	487 (36.7)
>5 cm	13 (4.7)	18 (6.0)	32 (6.1)	27 (12.1)	90 (6.8)
Unknown	5 (1.8)	8 (2.7)	0	6 (2.7)	19 (1.4)
Nodal status at diagnosis, n (%)					
Negative	151 (54.7)	142 (47.2)	352 (67.0)	132 (58.9)	777 (58.6)
Positive	124 (44.9)	158 (52.5)	167 (31.8)	91 (40.6)	540 (40.7)
Unknown	1 (0.4)	1 (0.3)	6 (1.1)	1 (0.4)	9 (0.7)
Tumor grade, n (%)					
1	34 (12.3)	35 (11.6)	48 (9.1)	23 (10.3)	140 (10.6)
2	137 (49.6)	153 (50.8)	368 (70.1)	128 (57.1)	786 (59.3)
3	100 (36.2)	111 (36.9)	104 (19.8)	73 (32.6)	388 (29.3)
Unknown	5 (1.8)	2 (0.7)	5 (1.0)	0	12 (0.9)
ER status, n (%)					
Positive	253 (91.7)	Not available	Not available	223 (99.6)	Not tabulated
Negative	23 (8.3)			1 (0.4)	
PR status, n (%)					
Positive	245 (88.8)	Not available	Not available	165 (73.7)	Not tabulated
Negative	31 (11.2)			59 (26.3)	
HER2 status, n (%)					
Positive	68 (24.6)	58 (19.3)	Not available	18 (8.0)	Not tabulated
Negative	202 (73.2)	136 (45.2)		154 (68.8)	

ARTICLE IN PRESS

ARTICLE IN PRESS

Equivocal	3 (1.1)	4 (1.3)		0 (-)	
Unknown	3 (1.1)	103 (34.2)		52 (23.2)	
Tamoxifen duration, years					
Median (range)	5.0 (4.7-5.4)	4.6 (2.7-5.0)	Not available	3.1 (2.0-5.0)	Not tabulated
Radiotherapy, n (%)					
Yes	182 (65.9)	251 (83.4)	Not available	Not available	Not tabulated
No	93 (33.7)	50 (16.6)			
Unknown	1 (0.4)	0			
Chemotherapy, n (%)					
Yes	225 (81.5)	223 (74.1)	Not available	79 (35.3)	Not tabulated
No	51 (18.5)	78 (25.9)		145 (64.7)	
TCF20 rs932376 genotype status, n (%)					
Wild-type	102 (37.0)	147 (48.8)	249 (47.4)	80 (35.7)	578 (43.6)
Heterozygous	130 (47.1)	120 (39.9)	217 (41.3)	123 (54.9)	590 (44.5)
Homozygous	44 (15.9)	34 (11.3)	59 (11.2)	21 (9.4)	158 (11.9)
CYP2D6 metabolizer status, n (%)					
Poor	0	23 (7.6)	45 (8.6)	23 (10.3)	91 (6.9)
Intermediate	127 (46.0)	126 (41.9)	197 (37.5)	118 (52.7)	568 (42.8)
Normal	143 (51.8)	149 (49.5)	270 (51.4)	70 (31.3)	632 (47.7)
Normal to Ultra-rapid	5 (1.8)	0	1 (0.2)	3 (1.3)	9 (0.7)
Ultra-rapid	1 (0.4)	3 (1.0)	12 (2.3)	10 (4.5)	26 (2.0)
Follow-up duration, years					
Median (range)	14.1 (10.3-16.0)	6.4 (5.1-7.7)	Not available	8.2 (6.5-9.8)	Not tabulated
Disease-free survival (DFS)					
No. of events / patients	42/276	93/301	116/525	59/224	310/1326
5-year DFS, % (95% CI)	95.3 (92.0-97.2)	78.2 (73.0-82.5)	82.6 (78.6-86.0)	80.6 (74.7-85.2)	83.9 (81.7-85.9)
Relapse Free Survival (RFS)					
No. of events / patients	32/276	86/301	89/525	43/224	250/1326
5-year RFS, % (95% CI)	95.3 (92.3-97.4)	78.5 (73.6-83.0)	86.3 (82.8-89.4)	86.0 (81.1-90.2)	86.3 (84.3-88.2)
Distant relapse-free survival (DRFS)					
No. of events / patients	20/276	67/301	60/525	38/224	185/1326
5-year DRFS, % (95% CI)	97.1 (94.6-98.6)	82.9 (78.4-87.0)	89.6 (86.5-92.3)	87.8 (83.1-91.7)	89.3 (87.5-91.0)

IQR, interquartile range; ER, estrogen receptor; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2

Note: Certain variables in the combined cohort were not computed as data was not available in at least one of the cohorts

ARTICLE IN PRESS

ARTICLE IN PRESS

List of Figures

Figure 1. Regional association plot for the chromosome 22 locus in the meta-analysis of breast cancer patients from Singapore, Germany, and Lebanon, between genetic markers and serum endoxifen concentration.

Genetic marker rs932376 shows the strongest evidence of statistical association in the locus, over and above that of markers mapping to CYP2D6. The horizontal axis reflects genes plotted according to chromosomal genomic position within the locus. The left vertical axis show P-values (-log10 scale) of the association between genetic markers and serum Z-endoxifen concentration. Genome-wide significance was prespecified at P < 5 ×10-8.

Figure 2. Forest plot depicting association of *TCF20* rs932376 per copy of minor allele and *CYP2D6* metabolizer status with endoxifen levels on multivariable analysis in combined discovery and validation cohorts

Figure 3. Comparison of (a) predictive values of *TCF20* rs932376 per copy of minor allele and *CYP2D6* metabolizer status on endoxifen levels, and (b) mean square errors of prediction models for endoxifen levels in combined discovery and validation cohorts.

In (a), *TCF20* rs932376 A>G was a significant predictor of endoxifen levels based on a likelihood ratio (LR) testing of a model with *CYP2D6* metabolizer status only with a model containing both genetic factors (*P*=0.009 in discovery and *P*=0.002 in validation). Similarly, *CYP2D6* metabolizer status was a significant predictor of endoxifen levels based on a LR testing of a model with *TCF20* rs932376 A>G only with a model containing both genetic factors (P=1.03X10^{-54} in discovery and P=6.92X10^{-17} in validation). However, *CYP2D6* metabolizer status accounted for a higher percentage of the predictive information contained in a bivariate model including both genetic factors as compared with *TCF20* rs932376 A>G (91.2% vs 48.8% in discovery 91.6% vs 56.1% in validation). In (b), the addition of *CYP2D6* metabolizer status to the model with only *TCF20* rs932376 A>G significantly improved the prediction accuracy of the resulting bivariate model with mean square errors (MSE) reduced from 497.5 to 451 (*P*<0.0001), while the addition of *TCF20* rs932376 A>G to the model with only *CYP2D6* metabolizer status did not significantly improve the prediction accuracy of the resulting bivariate model with MSE reduced marginally from 458.2 to 451 (*P*=0.0977).

Figure 4. Forest plot depicting association of TCF20 rs932376 per copy of minor allele and CYP2D6 metabolizer status with breast cancer outcomes in combined survival cohorts. DFS, disease free survival; RFS, relapse free survival; DRFS, distant relapse free survival; HR, hazard ratio; CI, confidence interval. A total of 53 patients (4%) with missing body mass index were excluded from the DFS and DRFS analyses.

List of Tables

Table 1. Patient demographics and clinical characteristics of discovery and validation cohorts used in genome-wide association study analysis

Table 2. Patient demographics, clinical characteristics and survival outcomes of cohorts used in survival analysis

ARTICLE IN PRESS