The Hospital for Sick Children

Technology Assessment at Sick Kids (TASK)

FULL REPORT

THIOPURINE S-METHYLTRANSFERASE TESTING FOR AVERTING DRUG TOXICITY IN PATIENTS RECEIVING THIOPURINES: A META-ANALYSIS OF DIAGNOSTIC TEST ACCURACY

Authors:

Richard M. Zur, Ph.D.
Research Project Manager, Child Health Evaluative Sciences,
The Hospital for Sick Children, Toronto

Lilla M. Roy, RN, BScN, M.Sc.
Research Project Coordinator, Child Health Evaluative Sciences,
The Hospital for Sick Children, Toronto

Wendy J. Ungar, MSc, PhD
Senior Scientist, Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto
Professor, Health Policy, Management & Evaluation, University of Toronto

Report No. 2015-03

October 7, 2015

Available at: http://lab.research.sickkids.ca/task/reports-theses/

Co-investigators:

Shinya Ito, MD, FRCPC
Division Head, Clinical Pharmacology and Toxicology, The Hospital for Sick Children, Professor, Medicine, Pharmacology & Pharmacy, Department of Paediatrics, University of Toronto

Elizabeth Uleryk, MLS
Director, The Hospital for Sick Children Library, Toronto

Joseph Beyene, MSc, PhD
Department of Clinical Epidemiology & Biostatistics, McMaster University

Knowledge user partners:

Chris Carew, MBA
Centre for Genetic Medicine, The Hospital for Sick Children

James Whitlock, MD
Division Head/Chief Haematology/Oncology, The Hospital for Sick Children; Professor, Paediatrics, University of Toronto

ACKNOWLEDGEMENTS

This research was supported by a Canadian Institutes of Health Research Knowledge Synthesis Grant, grant #305352. We thank Mr. Ian Schiller, M.Sc., Division of Clinical Epidemiology, McGill University Health Centre and Dr. Nandini Dendekuri, Ph.D., Director, Technology Assessment Unit of the McGill University Health Centre for their technical assistance with this report. We thank Ms. Christine Millan for administrative assistance.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

Correspondence:

Wendy J. Ungar, M.Sc., Ph.D.
Senior Scientist, Child Health Evaluative Sciences
The Hospital for Sick Children Peter Gilgan Centre for Research and Learning
11th floor, 686 Bay Street
Toronto, ON, Canada M5G 0A4
tel: (416) 813-7654, extension 303487
fax: (416) 813-5979
e-mail: wendy.ungar@sickkids.ca
http://www.sickkids.ca/AboutSickKids/Directory/People/U/Wendy-Ungar.html

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

ABBREVIATIONS

6-MMP	6-methyl-mercaptopurine
6-MTG	6-methylthioguanine
ADE	Adverse drug event
AHRQ	Agency for Healthcare Research and Quality
ALL	Acute lymphoblastic leukemia
ARMS	Multiplex amplification refractory mutation
AS-PCR	Allele-specific polymerase chain reaction
CCTR	Cochrane Central Register of Controlled Trials
CDSR	Cochrane Database of Systematic Reviews
CEA	Cost-effectiveness analysis
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CI	Confidence interval
CrI	Credible interval
DARE	Database of Abstracts of Reviews of Effects
DHPLC	Denaturing high performance liquid chromatography
DNA	Deoxyribonucleic acid
DTA	Diagnostic test accuracy
gHb	Gram of hemoglobin
h	Hour
Hb	Hemoglobin
HPLC	High performance liquid chromatography
HSROC	Hierarchical summary receiver operating characteristic
HTA	Health Technology Assessment
IBD	Inflammatory bowel disease
IPA	International Pharmaceutical Abstracts
Mg	Milligram
Min	Minute
mL	Millilitre

MTG	Methylthioguanine
NHSEED	National Health Service Economic Evaluation Database
Nmol	Nanomole
Pmol	Picomole
PCR	Polymerase chain reaction
pRBC	Packed red blood cells
QUADAS-2	Quality Assessment Tool for Diagnostic Accuracy Studies
RBC	Red blood cell
RC	Radiochemical method
RFLP	Restriction fragment length polymorphism
ROC	Receiver operating characteristic
SROC	Summary receiver operating characteristic
SSCP	Single strand conformation polymorphism
TPMT	Thiopurine S-methyltransferase
U	Unit

EXECUTIVE SUMMARY

Introduction

Thiopurine S-methyltransferase (TPMT) is an enzyme that metabolizes thiopurine drugs. The absence or a deficiency in TPMT activity can significantly increase the risk of an adverse drug event (ADE) in persons receiving thiopurine therapy as they are unable to properly metabolize the drug. Unless thiopurine drugs are avoided or doses are reduced in these patients, they are at greater risk for life-threatening bone marrow toxicity and liver toxicity, which may lead to myelosuppression, anemia, bleeding, leukopenia, infection, and death. There are two approaches to testing for TPMT deficiency. Phenotype tests that measure levels of TPMT enzyme activity *in vitro* are common. Alternatively, genotype tests are available that detect the presence of variants in the genes responsible for expressing the TPMT enzyme. It remains uncertain whether an enzyme activity (phenotype) or genotype diagnostic test is the most appropriate strategy for clinical practice. Numerous studies have been performed to assess the accuracy of both types of diagnostic tests, however meta-analyses that summarize all available evidence have been limited due to the technical challenges with pooling diagnostic test accuracy (DTA) results and the lack of a gold reference standard.

Objectives

The aim of this study was to meta-analyze the sensitivity and specificity of phenotype and genotype TPMT testing reported in the literature. The specific objectives were:

1. To perform meta-analyses of two methods of evaluating TMPT enzyme activity: a) identifying patients with deficient or absent TPMT enzyme activity (patients that are homozygous for TPMT mutations) versus the rest of the population and b) identifying patients that have either low or intermediate TPMT enzyme activity (patients that are homozygous or heterozygous for TPMT mutations) versus the rest of the population.

2. To perform a DTA meta-analysis that accounts for the imperfect reference standard provided by genotype testing.

Methods

A comprehensive systematic review and critical appraisal of all published studies of TPMT test accuracy were conducted in the first phase of this research. Two different testing approaches were considered: 1) tests identifying patients with deficient or absent TPMT enzyme activity (patients that are homozygous for TPMT mutations) versus the rest of the population and 2) tests identifying patients that have either low or intermediate TPMT enzyme activity (patients that are homozygous or heterozygous for TPMT mutations) versus the rest of the population. The meta-analysis was performed using a hierarchical summary receiver operating characteristic (HSROC) approach. A latent class meta-analysis method that allowed for heterogeneity in cut-point definition in phenotype TPMT testing while also allowing for an imperfect reference standard was used to meta-analyze the sensitivity and specificity data for the two approaches.

Results

When identifying patients with deficient or absent TPMT enzyme activity (patients that are homozygous for TPMT mutations), the latent class model resulted in a pooled sensitivity and specificity of phenotype testing of 75.9% (95% credible interval [CrI], 58.3% to 87.0%) and 98.9% (95% CrI, 96.3% to 100%), respectively. The latent class meta-analysis also provided pooled sensitivity and specificity of the genotype tests. For genotype tests evaluating only the most common TPMT*2 and TPMT*3 polymorphisms, the pooled sensitivity and specificity was 90.4% (95% CrI, 79.1% to 99.4%) and 100.0% (95% CrI, 99.9% to 100%), respectively. For genotype tests evaluating TPMT*2, TPMT*3 and more polymorphisms, the pooled sensitivity

and specificity was 80.7% (95% CrI, 41.7% to 99.4%) and 99.9% (95% CrI, 99.7% to 100%), respectively.

When testing individuals to detect deficient or intermediate TPMT activity (homozygous or heterozygous TPMT mutations) versus the remainder of the population, the pooled sensitivity and specificity of phenotype testing was 91.3% (95% CrI, 86.4% to 95.5%) and 92.6% (95% CrI, 86.5% to 96.6%), respectively. For genotype tests evaluating TPMT*3 mutations only, the pooled sensitivity and specificity was 66.8% (95% CrI, 51.1% to 94.6%) and 99.9% (95% CrI, 99.5% to 100%), respectively. For genotype tests evaluating TPMT*2 and TPMT*3 only, the pooled sensitivity and specificity was 88.9% (95% CrI, 81.6% to 97.5%) and 99.2% (95% CrI, 98.4% to 99.9%), respectively. For genotype tests evaluating TPMT*2, TPMT*3, and more polymorphisms, the pooled sensitivity and specificity was 93.5% (95% CrI, 84.9% to 99.3%) and 99.9% (95% CrI, 99.7% to 100%), respectively.

Conclusions

The pooled estimates of sensitivity suggest that genotype testing has higher sensitivity than phenotype testing as long as both TPMT*2 and TPMT*3 polymorphisms are tested. However, due to the large 95% CrIs around sensitivity estimates the results are not statistically significant. Both tests have been shown to have high specificity, valuable for ruling in the presence of TPMT deficiency. This meta-analysis cannot conclude that one test is superior to the other. Although more complex than standard meta-analysis techniques, the latent class HSROC approach is straight-forward to implement and interpret. Therefore, this report supports existing recommendations to perform HSROC or bivariate methods for DTA meta-analyses.

1 INTRODUCTION

The present meta-analysis represents the second phase of a program of research to synthesize the evidence regarding thiopurine S-methyltransferase (TPMT) pharmacogenetic testing. A comprehensive systematic review and critical appraisal of all published studies of TPMT test accuracy were completed for the first phase of this research (1).

1.1 Background

Thiopurine S-methyltransferase (TPMT) is an enzyme that metabolizes thiopurine drugs. Thiopurines are commonly used in maintenance treatment for childhood leukemias, as well as, less commonly, for inflammatory bowel disease, transplant recipients, and dermatological conditions. The absence or a deficiency in TPMT activity can significantly increase the risk of an adverse drug event (ADE) in persons receiving thiopurine therapy as they are unable to properly metabolize the drug. Unless thiopurine drugs are avoided or doses are reduced in these patients, they are at greater risk for life-threatening bone marrow toxicity and liver toxicity, which may lead to myelosuppression, anemia, bleeding, leukopenia, infection, and death (2).

There are two approaches to testing for TPMT deficiency. Phenotype tests that measure levels of TPMT enzyme activity *in vitro* are common. Alternatively, genotype tests are available that detect the presence of variants in the genes responsible for expressing the TPMT enzyme (3). Both tests have associated challenges and it remains uncertain whether an enzyme activity (phenotype) or genotype diagnostic test is the most appropriate strategy for clinical practice.

Phenotype test results can be confounded by concomitant medications or blood transfusions (4-11). In addition, phenotype tests require specification of a cut-point to establish a positive test result. The choice of cut-point is crucial in phenotype tests. Typically, two cut points are required: a low cut-point (e.g. 5 U per mL packed red blood cells) is used for identifying patients with deficient TPMT activity in whom thiopurines should be avoided, and an intermediate cut-point (e.g., 15 U per mL packed red blood cells) is used for identifying patients with reduced activity for whom a reduced dose would be safer. The choice of cut-point is influenced by the patient population and the clinical indication for testing, as it may be more apt to risk an ADE for a greater chance of successful treatment for certain patient populations (12).

Genotype tests are often limited by the number of genes that can be tested simultaneously (either due to physical or cost constraints). While there are 24 genes implicated in TPMT, 3 variants (*3A, *14A and *22) account for 90% of the deficiencies occurring in the population. (13). Patients with these three variants have no detectable enzyme activity. Patients with other variants have approximately 50% of functional enzyme activity (13). The most common TPMT genomic tests include only TPMT*2 and TPMT*3, and as a result leave patients with rare mutations at risk (13). The prevalence of mutations is known to vary by ethnic background (8, 14-16), thus certain segments of the population may be more at risk.

1.2 Previous studies

Several systematic reviews have been undertaken to examine the clinical validity of TPMT phenotype and genotype tests. A recent systematic review by our group found that there is a large literature of studies evaluating TPMT phenotype and genotype testing, but in general studies reported a wide range of test performance characteristics and did not focus on the evaluation of diagnostic test accuracy (17, 18).

Previous studies of the sensitivity and specificity of phenotype and genotype TPMT testing have used different approaches to derive a summary estimate. In our group's economic evaluation of

TPMT testing for children with acute lymphoblastic leukemia (ALL), midpoint rather than pooled estimates of sensitivity and specificity were used (6, 19). In a meta-analysis performed by the US Agency for Healthcare Research and Quality (AHRQ), sensitivity and specificity estimates were transformed first to make them more normally distributed before the independent mean estimates were calculated (17). However, that study did not address the correlation between sensitivity and specificity. Thus there are gaps in the evidence with regard to the methods for pooling the performance characteristics of diagnostic tests in general, and for TPMT tests in particular.

1.3 Meta-analysis of diagnostic test accuracy

A diagnostic test's accuracy can be summarized with two parameters: the sensitivity and specificity. The sensitivity is the proportion of cases with the disorder of interest correctly classified and the specificity is the proportion of cases without the disorder of interest correctly classified. If the test has a variable cut-point for defining a tested case as positive or negative, the test's sensitivity and specificity will vary together as the cut-point is changed. Receiver operating characteristic (ROC) analysis summarizes the inherent trade-offs between sensitivity and specificity as the decision threshold is made more or less stringent, and is an established methodology for the assessment of diagnostic performance.

Determining the performance characteristics of TPMT phenotype and genotype testing would be aided through a meta-analysis of published findings. An accurate pooled estimate of performance characteristics is also needed for assessments of cost-effectiveness. However, three difficulties exist in meta-analysing phenotype and genotype TPMT test performance results. First, it is well-known that the meta-analysis of diagnostic test performance is more complicated than meta-analysis of treatment effects from randomized controlled trials (20-23).

Measures of diagnostic accuracy such as sensitivity and specificity are correlated and sophisticated statistical analysis is required to account for that correlation (24-27). Failure to account for that correlation may result in inaccurate pooled point estimates. Sensitivity and specificity should be pooled simultaneously, taking account of the correlation. Secondly, it is natural to have heterogeneity in the data when the choice of cut-point for deciding between normal TPMT enzyme activity and low enzyme activity varies between studies and patient populations. If the cut-point varies between studies the resulting heterogeneity amongst the sensitivity and specificity pairs makes it unreasonable to simply pool the results into a single sensitivity and specificity pair. Ignoring the variable nature of the cut-point could lead clinicians to falsely conclude that two tests are different when they actually have the same discriminatory ability. Statistical models exist that attempt to account for varying cut-points between studies, but they have been challenging to use as they rely on computationally-intense Bayesian models (24, 26, 27). Third, neither the phenotype test nor the genotype test is an appropriate gold standard (28-30). A gold standard would be an independent test that is very likely to accurately classify the case as positive or negative, and the index test (i.e., phenotype or genotype) could be compared to that gold standard to determine their accuracy. Without a gold standard, the phenotype test and genotype tests can only be compared to each other, in other words, utilizing an imperfect gold standard. Therefore, a meta-analysis should either assume a missing gold standard or that the method chosen as the reference standard, whether it is the phenotype test or genotype test, is an imperfect gold standard.

Two methods exist that address these challenges associated with meta-analysis of diagnostic tests: the bivariate method (31) and the hierarchical summary ROC curve method (HSROC) (26). When no meta-regression is performed it has been shown that both methods provide mathematically equivalent results (32, 33). The difference is that the bivariate method provides summary sensitivity and specificity pairs whereas the HSROC method is parameterized in terms of accuracy and threshold, and generates a hierarchical summary ROC curve that best fits a group of sensitivity and specificity pairs.

Therefore, diagnostic test accuracy (DTA) meta-analyses must address the issues of expected heterogeneity as well as correlation of study parameters. A DTA for phenotype and genotype testing of TPMT enzyme activity must also address the issue of imperfect reference standards.

1.4 Objectives

The aim of this study was to meta-analyze the sensitivity and specificity of phenotype and genotype TPMT testing reported in the literature. The specific objectives were:

1. To perform meta-analyses of two methods of evaluating TMPT enzyme activity: a) identifying patients with deficient or absent TPMT enzyme activity (patients that are homozygous for TPMT mutations) versus the rest of the population and b) identifying patients that have either low or intermediate TPMT enzyme activity (patients that are homozygous or heterozygous for TPMT mutations) versus the rest of the population.
2. To perform a DTA meta-analysis that accounts for the imperfect reference standard provided by genotype testing.

2 METHODS

2.1 Study design

A comprehensive systematic review of all published studies of TPMT test performance was initially conducted (1). Subsequently, a meta-analysis of the sensitivity and specificity of phenotype testing and genotype testing for TPMT was undertaken. A latent class meta-analysis method that allowed for heterogeneity in cut-point definition in phenotype TPMT testing while also allowing for an imperfect reference standard was used.

Two different testing approaches were considered: 1). identifying patients with deficient or absent TPMT enzyme activity (patients that are homozygous for TPMT mutations) versus the rest of the population and 2) identifying patients that have either low or intermediate TPMT enzyme activity (patients that are homozygous or heterozygous for TPMT mutations) versus the rest of the population. The first of these is the more clinically relevant test, as patients with deficiency or homozygous for mutations are at much greater risk of ADE and will have their drug regimen changed to avoid complications related to thiopurine use.

2.2 Data sources and searches

Electronic citation databases and grey literature sources were searched for relevant publications. The search included the following databases: Biosis, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Cochrane Database of Systematic Reviews (CDSR), Cochrane Central Register of Controlled Trials (CCTR), Database of Abstracts of Reviews of Effects (DARE), Health Technology Assessment (HTA), National Health Service Economic Evaluation Database (NHSEED), Embase, International Pharmaceutical Abstracts (IPA), Medline, and PubMed. Studies in any language comparing a genotype or phenotype technology to another genotype or phenotype technology were included. Studies must have been conducted in humans, and must have reported sufficient data to calculate sensitivity, specificity, negative predictive value, positive predictive value or concordance between the two technologies.

Grey literature was obtained directly from web sites of government health agencies, health technology assessment agencies and institutions, health economic research groups, research institutes, academic organizations such as universities, and websites related to the diseases of interest (e.g. ALL). Detail describing the search can be found elsewhere (1).

2.3 Study selection

Eligible studies were those that: 1) evaluated either a TPMT genotype or TPMT phenotype technology in comparison to a reference standard; 2) presented results on the accuracy of the two tests, using either sensitivity and specificity, or positive/negative predictive values together with prevalence, or presented raw data in the text, in supplemental files, or directly from the study authors to allow these measures to be calculated; 3) were conducted in any age group; 4) were conducted in any disease group; and 5) were published in any language, so long as it was possible to obtain sufficient translation to determine eligibility. Studies not conducted in humans, including animal, tissue and *in vitro* studies were excluded.

Two reviewers (R.Z. and L.R.) performed the screening and selection of studies. Initially reviewers independently reviewed titles and abstracts for inclusion according to the previously described criteria. All abstracts and titles were categorized for eligibility as ‘yes’, ‘no’, or ‘maybe’. The categorization was compared between reviewers after approximately 60 titles and abstracts. Discrepancies were resolved by establishing a set of decision rules, in consultation with the principal investigator (W.U.) as needed. Agreement became consistent after comparing categorization of approximately 130 abstracts and titles between the two reviewers. Subsequently, one reviewer (L.R.) screened the remaining titles and abstracts.

2.4 Qualify assessment and data extraction

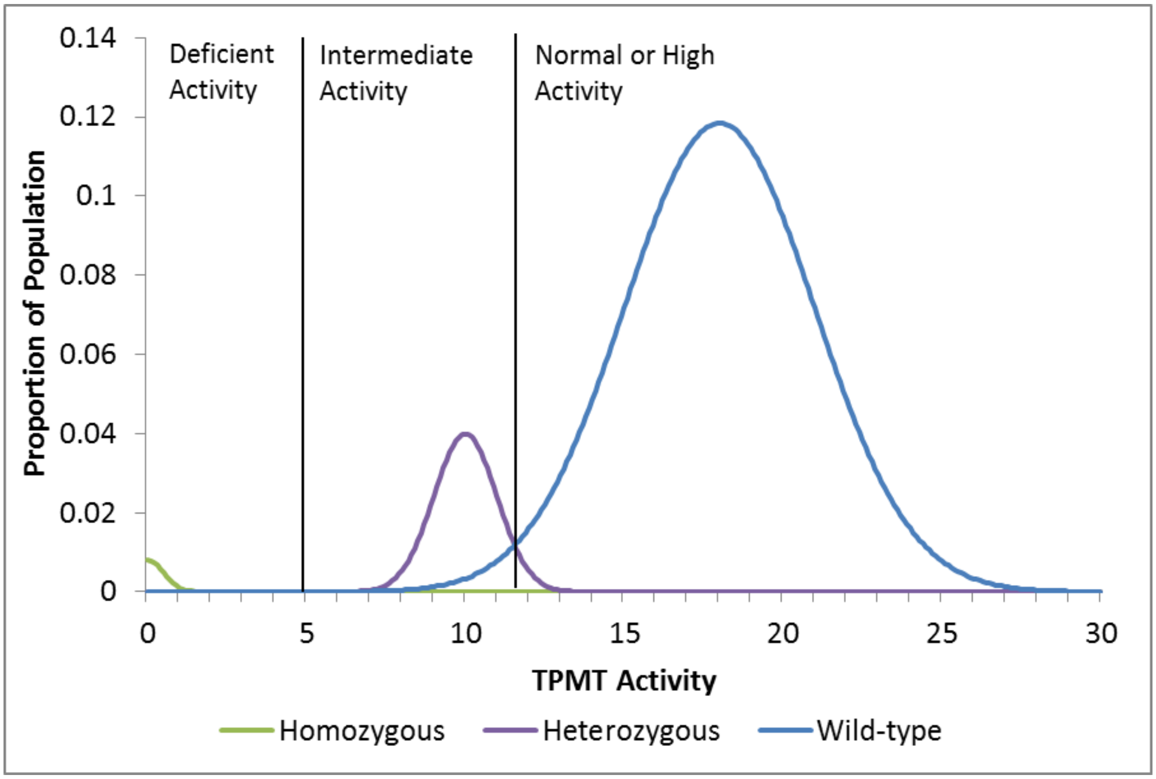
A quality appraisal of eligible publications was conducted using the QUADAS-2 (Quality Assessment Tool for Diagnostic Accuracy Studies) (34). Signaling questions for the QUADAS-2 tool were developed that were relevant for assessing the quality of genetic diagnostic tests. Two authors (L.R. and R.Z.) carried out quality assessment and data extraction independently on 5% of the articles and discussed any discrepancies to arrive at a consensus assessment. A single author (L.R.) performed quality assessment and data extraction on the remaining articles after consensus was reached on the initial 5%. If a remaining article was difficult to assess it was also discussed and consensus was reached.

An article was judged to be high quality if all five QUADAS domains demonstrated low bias and had low concern for applicability. If all of the domains were unclear or had high risk of bias, then the study was judged to be low quality. If only one domain demonstrated high risk of bias, then the study was judged to be of high quality. If the study had two or more domains that were uncertain, then the study was judged to be low quality.

Data from each included study were extracted into a custom-made MS Access database. Data extraction included basic study design characteristics, study results, diagnostic test performance characteristics, and data required to populate 2x2 or 3x3 contingency tables for the calculation of sensitivity and specificity.

Because of the nature of TPMT activity in patients, articles typically categorized data into three categories: deficient, intermediate, or normal/high for phenotype testing and homozygous, heterozygous, or wild-type for genotype tests. An example of the distribution of patients within a population is shown in Fig. 1, together with approximate cut-points for the phenotype test. However, ROC analysis can only address binomial tests, so even though 3x3 contingency tables were available for some studies, 2x2 contingency tables were estimated for analysis.

Figure 1. Illustration of TPMT mutation and activity distribution



2.5 Meta-analysis

A latent class meta-analysis method that allowed for heterogeneity in cut-point definition in phenotype TPMT testing while also allowing for an imperfect reference standard was used (29). It is known that the phenotype test has a cut-point based on enzyme activity that can be varied whereas the genotype test does not. Although there was not extensive heterogeneity, the phenotype tests displayed a variety of cut-points throughout the various studies. A random effects model that accounted for between-test variability was therefore justified. Therefore, it is natural to estimate the SROC curve for the phenotype test and sensitivity and specificity values for the genotype test. The statistical model simultaneously estimated the SROC curve for the phenotype test and sensitivity and specificity values for the genotype test.

The model assumed that the data were conditionally independent. In other words, the model assumed that genotype test results were unrelated to phenotype test results for each case. A conditionally dependent model is available (27, 35), but it is not well-specified and it was found that the program used to fit the model would consistently fail to fit the data (36). Therefore the conditionally independent model was used for this meta-analysis.

The meta-analysis model can account for multiple different reference tests. Because the sensitivity and specificity of the genotype test varies with the polymorphisms tested, studies were stratified based on the polymorphisms included in the test. This resulted in two groups for the test of deficient TPMT enzyme activity or homozygous TPMT mutation: 1) studies where the genotype test only considered TPMT*2 and *3 (37-46), and 2) studies where the genotype test considered TPMT*2, TPMT*3 and additional polymorphisms (47-49).

There were three reference groups for the test of low or intermediate TPMT enzyme activity or homozygous or heterozygous TPMT mutation: 1) studies where the genotype test only considered TPMT*3 (50, 51), 2) studies where the genotype test considered TPMT*2 and TPMT*3 (10, 37-46, 52-60), and 3) studies where the genotype test considered TPMT*2, TPMT*3 and additional polymorphisms (46-49, 61, 62).

The systematic review included studies that reported sensitivity, specificity, positive predictive values, negative predictive values, and prevalence. For the meta-analysis, only studies that allowed construction of 2x2 contingency tables for calculation of sensitivity and specificity from the published results were included.

The meta-analysis was a Bayesian model. Non-informative prior distributions that allowed the data to dominate the final estimate of the SROC curve and sensitivity and specificity values were used. Factors for meta-regression were not included because the number of studies was not large enough. Estimates of the mean and 95% credible intervals (CrI) of the pooled sensitivity and specificities for the phenotype TPMT test across all studies, and for the genotype TPMT test pooled by group based on the number of polymorphisms tested, were derived. An SROC curve was estimated for phenotype TPMT testing. Analyses were carried out in WinBUGS version 1.4.3 and R version 3.0.2. The WinBUGS programs were obtained online (63).

The results are presented in forest plots, where all included studies are listed, in alphabetical order, and their associated data and sensitivity and specificity values are plotted with associated 95% confidence intervals. The result of the meta-analysis is an SROC curve overlaid onto the studies’ sensitivity and specificity pairs. A pooled sensitivity and specificity estimate is provided on the SROC curve with an associated 95% credible region.

3 RESULTS

3.1 Search results

Through database and grey literature searching 4071 potentially eligible studies were identified. Of the identified studies, 373 records required full text review and 121 records were identified for closer review to establish whether they contained relevant data for extraction. Ultimately, 66 studies were identified as having sufficient data for inclusion, and underwent quality appraisal. The search results are summarized in the PRISMA diagram in the Appendix.

3.2 Quality assessment

The 66 studies with sufficient data for inclusion comprised three categories: 1) phenotype-genotype comparisons, 2) phenotype-phenotype comparisons, and 3) genotype-genotype comparisons. In total, 55 studies contained phenotype-genotype comparisons and were evaluated for inclusion in the meta-analysis.

Of the 55 studies, 30 were designated as high quality by the quality appraisal. All of the high quality studies were published between 1997 and 2013, and examined a range of genotype and phenotype test methods. Of the 30 high-quality studies, 27 provided sufficient data for the analysis of deficient or intermediate TPMT activity versus the remainder of the population, and 13 studies provided sufficient for the analysis of deficient TPMT activity versus the remainder of the population.

The ethnic population studied, the population's disease, the amplification/genotype method and polymorphisms tested, together with the phenotype method, phenotype cutpoints, and units of measurement of the phenotype test are presented for each study in Table 1. The most common test populations included healthy individuals, as well as patients with inflammatory bowel disease or acute lymphoblastic leukemia. The primary type of amplification method was polymerase chain reaction (PCR), and the majority of genotype tests used allele-specific polymerase chain reaction (AS-PCR) or restriction fragment length polymorphism (RFLP). Most studies tested TPMT*2 and TPMT*3 polymorphisms. The majority of phenotype tests used radiochemical method (RC) or high performance liquid chromatography (HPLC) testing and defined cut-points to classify patients as either deficient, low activity, or normal/high activity.

Table 1. Genotype test characteristics and polymorphisms tested in phenotype-genotype studies

Author	Year	Population	Disease	Amplification/ Genotype Method	Polymorphisms Tested	Phenotype Method	Cutpoints	Unit
Ben Salah	2013	Other	Crohn's disease	PCR; AS-PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	HPLC	Low (not specified), intermediate (5-10), high (>10),	nmol 6-MMP/h/ml pRBC
Fakhoury	2007	European	Acute lymphoblastic leukemia	PCR; AS-PCR	TPMT*2, TPMT*3a, TPMT*3c	HPLC	Intermediate (<11.8); deficient estimated from graph as approximately 6	U/mL pRBCs
Fangbin	2012	Chinese	Inflammatory bowel disease	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	Not specified	Optimal cutoff calculated by ROC: intermediate (<4.75) (heterozygous carrier).	U/mL RBC
Ford	2006	Not specified	Not specified	ARMS; AS-PCR; PCR	TPMT*2, TPMT*3	HPLC	Researchers calculated own cutpoint for low/intermediate; unclear whether they calculated it for high	nmol 6-MTG/gHb/h
Ganiere-Monteil	2004	Caucasian	Otherwise healthy	PCR; AS-PCR	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	HPLC	Post-hoc suggestion of phenotype cut-off (13.5) between wild-type and heterozygous genotype.	U/mL pRBC

Author	Year	Population	Disease	Amplification/ Genotype Method	Polymorphisms Tested	Phenotype Method	Cutpoints	Unit
Gazouli	2012	Not specified	Inflammatory bowel disease	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	RC	Low (<5.5), intermediate (5.6-15.5); normal-high (>15.6)	U/mL RBC
Hindorf	2012	Not specified	Inflammatory bowel disease	Pyrosequencing	TPMT*2, TPMT*3a, TPMT*3c; those with phenotype under 9.0 were further investigated on exons 3-10.	RC	Low (<2.5); high (>9.0)	U/mL pRBC
Jorquera	2012	Other	Otherwise healthy	PCR; RFLP	TPMT*1, TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	HPLC	Deficient (</=5); low (6-24); normal (25-55); high (>/=56)	nmol/gHb/h
Langley	2002	Not specified	Autoimmune liver disease - (autoimmune hepatitis)	PCR; RFLP	TPMT*3a, TPMT*3b, TPMT*3c	RC	Deficient (<5.0); intermediate (5-13.7); high (>13.7)	U/ml
Larussa	2012	Caucasian	Inflammatory bowel disease	PCR; RFLP	TPMT*2, TPMT*3b, TPMT*3c	Competitive micro-well immunoassay	Very low (</=5.5); intermediate (5.6-15.5); normal to hi (>/=15.6)	U/gHb
Lennard	2012	Other	Acute lymphoblastic leukemia	PCR; RFLP	TPMT*3a, TPMT*3b, TPMT*3c	HPLC	Between intermediate and high - varied cutpoints at 9.5, 10.5, 11.5	Units/mL pRBC
Liang	2013	Not specified	Organ transplant	PCR; TaqMan	TPMT*2, TPMT*3a, TPMT*3c	Not specified	Low (<6.3); intermediate (6.3-15.0); normal	U/ml RBC

Author	Year	Population	Disease	Amplification/ Genotype Method	Polymorphisms Tested	Phenotype Method	Cutpoints	Unit
							(15.1-26.4)	
Loennechen	2001	Caucasian	Patients admitted to a cardiology centre	PCR; AS-PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*6	RC	Deficient (<5); heterozygous intermediate (5-9.5); wild-type (>9.5)	U/mL pRBC
Ma	2006	Chinese	Acute lymphoblastic leukemia	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3c	HPLC	12	U
Marinaki	2003	Caucasian	Inflammatory bowel disease and dermatology patients	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3c	RC	Low (<2.5); intermediate (2.5-8); normal (8-15)	nmol 6-MMP/h/ml RBC
Milek	2006	Other	Otherwise healthy	PCR; RFLP, TaqMan	TPMT*2, TPMT*3b, TPMT*3c	HPLC	Calculated using ROC analysis; Low<5.8 assumed based on previous study as not reported in this study; High >9.82	pmol 6-MMP/ 10^7 RBC/h
Schaeffeler	2004	Caucasian	Otherwise healthy	PCR; DHPLC	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D	RC	Low (<9), intermediate (9-22); high (22-50); very high (51-65)	mmol
Schwab	2002	Caucasian	Inflammatory bowel disease	DHPLC	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D	Not specified	High (>24); low (<3)	nmol 6-MTG/ gHb /h

Author	Year	Population	Disease	Amplification/ Genotype Method	Polymorphisms Tested	Phenotype Method	Cutpoints	Unit
Serpe	2009	Other	Otherwise healthy	AS-PCR; PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	Not specified	"arbitrary cutpoints" low (<8.0); intermediate (<19.4); normal (<37.0); high (>37.0)	U/gHb; nmol 6-MMP/h
Spire-Vayon de la Moureyre	1998	European	Otherwise healthy	PCR-SSCP; Direct sequencing	TPMT*1, TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*1S, TPMT*1A, TPMT*7, TPMT *3d	RC	Deficient (<5 U/ml); intermediate (5-13.7), high (>13.7),	U/ml RBC
Spire-Vayron de la Moureyre	1998	European	Not specified	PCR-SSCP; Direct sequencing	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D, TPMT*4, TPMT*5, TPMT*6, TPMT*7	RC	Low (<5); intermediate (5-13.7); high (>13.7)	U/mL RBC
von Ahsen	2005	Caucasian	Inflammatory bowel disease	Not specified	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	RC	Low (<10)	nmol/(mL RBC/h)
Wennerstrand	2013	Other	Acute lymphoblastic leukemia	Pyrosequencing	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D	RC	Low vs intermediate (2.5); high vs intermediate (9.0)	U/mL pRBC

Author	Year	Population	Disease	Amplification/ Genotype Method	Polymorphisms Tested	Phenotype Method	Cutpoints	Unit
Winter	2007		Inflammatory bowel disease or ulcerative colitis	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	Mass spectrometry	Low (<10), intermediate (10-25); normal (26-50); high (>50)	pmol/h/mg Hb
Xin	2009	Not specified	Organ transplant	AS-PCR; PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	HPLC	Very low (<3); intermediate (3-24); normal (24-50); high (>50U)	U
Yates	1997	Caucasian	Acute lymphoblastic leukemia	PCR; RFLP	TPMT*1, TPMT*2, TPMT*3a, TPMT*3c	RC	Deficient (<5.0); heterozygous (5-10); homozygous wild-type (>10)	U/ml pRBC
Zhang	2007	Not specified	Chronic renal failure	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c	HPLC	Calculated by ROC	nmol/ml pRBC

Abbreviations: ARMS (multiplex amplification refractory mutation); AS-PCR (allele-specific polymerase chain reaction); DHPLC (denaturing high performance liquid chromatography); gHb (Gram of hemoglobin); h (hour); Hb (hemoglobin); HPLC (high performance liquid chromatography); mg (milligram); ml (milliliter); MTG (methylthioguanine); nmol (nanomole); PCR (polymerase chain reaction); pmol (picomole); pRBC (packed red blood cells); RBC (red blood cell); RFLP (restriction fragment length polymorphism); SSCP (single strand conformational polymorphism); TPMT (thiopurine s-methyltransferase); U (unit); 6-MMP (6-methyl-mercaptopurine); 6-MTG (6-methylthioguanine)

3.3 Meta-analysis

Figure 2 summarizes the data related to 13 studies that assessed testing for deficient TPMT enzyme activity or homozygous TPMT mutations. Ranges for sensitivity and specificity of the genotype test reference standard were 50-100% and 88-100%, respectively. The prevalence of deficient TPMT enzyme activity or homozygous TPMT mutations in the study sample ranged from 0.2% to 14.2%.

The forest plots in Figure 2 show 11 out of 13 studies with perfect sensitivity. However, 8 of those studies had 95% confidence intervals (CIs) that covered nearly the full range of possible sensitivity values. This is a result of the very low prevalence of deficient TPMT activity, as 8 out of 13 studies had only 1 or 2 cases with deficient TPMT activity. The specificities are also estimated to be very large, with 11 out of 13 studies having perfect specificity. However, the 95% CIs for specificities are very narrow.

Figure 3 summarizes the data from the 27 studies that assessed testing to identify cases of deficient or intermediate TPMT enzyme activity (higher cut-point). Ranges for sensitivity and specificity of the genotype test reference standard were 60-100% and 60-100%, respectively. The prevalence of deficient or intermediate TPMT enzyme activity ranged from 2.3% to 55.7%.

Figure 2. Forest plot of sensitivities and specificities of phenotype versus genotype tests discriminating deficient TPMT individuals versus others

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ben Salah, L., et al., 2013	1	0	0	87	1.00 [0.03, 1.00]	1.00 [0.96, 1.00]
Fakhoury, M., et al., 2007	2	0	0	118	1.00 [0.16, 1.00]	1.00 [0.97, 1.00]
Ford, L., Graham, V., Berg, J., 2006	2	0	0	400	1.00 [0.16, 1.00]	1.00 [0.99, 1.00]
Ganiere-Monteil, C., et al., 2004	1	0	0	467	1.00 [0.03, 1.00]	1.00 [0.99, 1.00]
Gazouli, M., et al., 2012	6	12	1	89	0.86 [0.42, 1.00]	0.88 [0.80, 0.94]
Hindorf, U., Appell, M.L., 2012	37	6	0	7152	1.00 [0.91, 1.00]	1.00 [1.00, 1.00]
Larussa, T., et al., 2012	1	5	1	44	0.50 [0.01, 0.99]	0.90 [0.78, 0.97]
Schaeffeler, E. et al., 2004	7	0	0	1207	1.00 [0.59, 1.00]	1.00 [1.00, 1.00]
Schwab, M., et al., 2002	1	0	0	92	1.00 [0.03, 1.00]	1.00 [0.96, 1.00]
Serpe, L., et al., 2009	1	0	0	942	1.00 [0.03, 1.00]	1.00 [1.00, 1.00]
Spire-Vayron de la Moureyre, C., et al., 1998a	1	0	0	34	1.00 [0.03, 1.00]	1.00 [0.90, 1.00]
Spire-Vayron de la Moureyre, C., et al., 1998b	1	0	0	190	1.00 [0.03, 1.00]	1.00 [0.98, 1.00]
Yates, C.R., et al., 1997	6	0	0	42	1.00 [0.54, 1.00]	1.00 [0.92, 1.00]

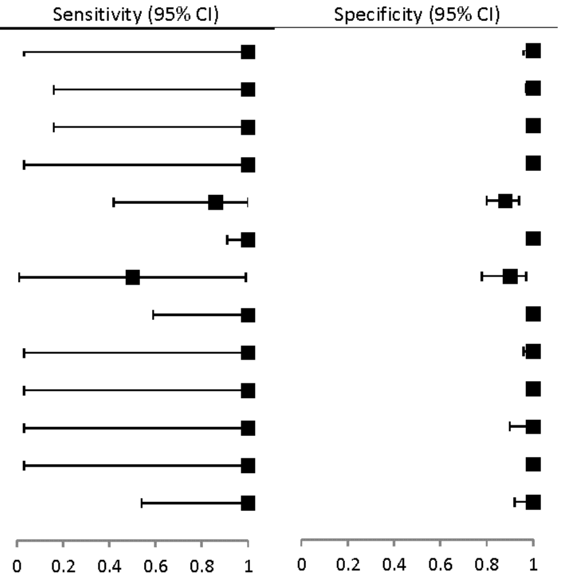
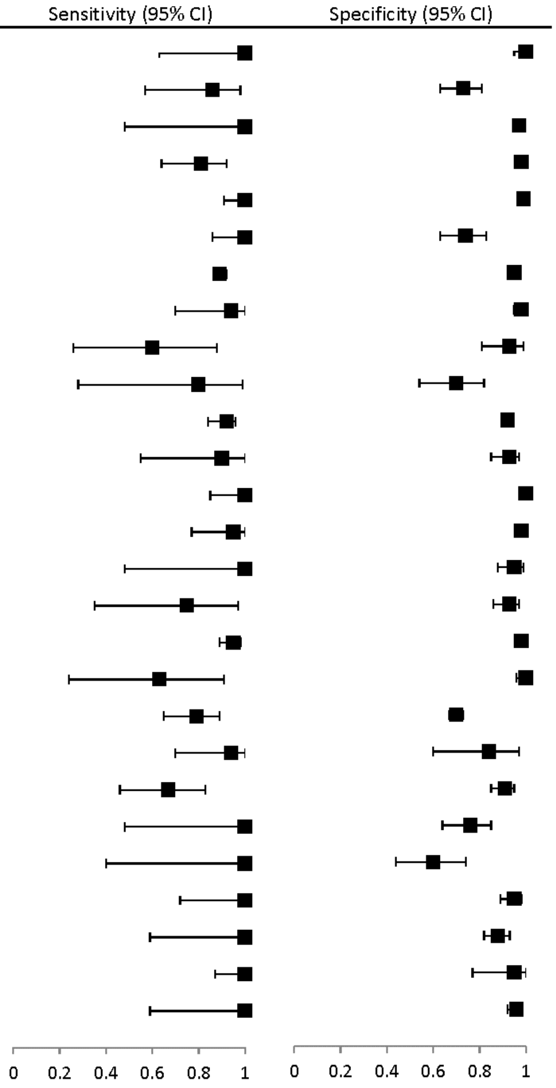


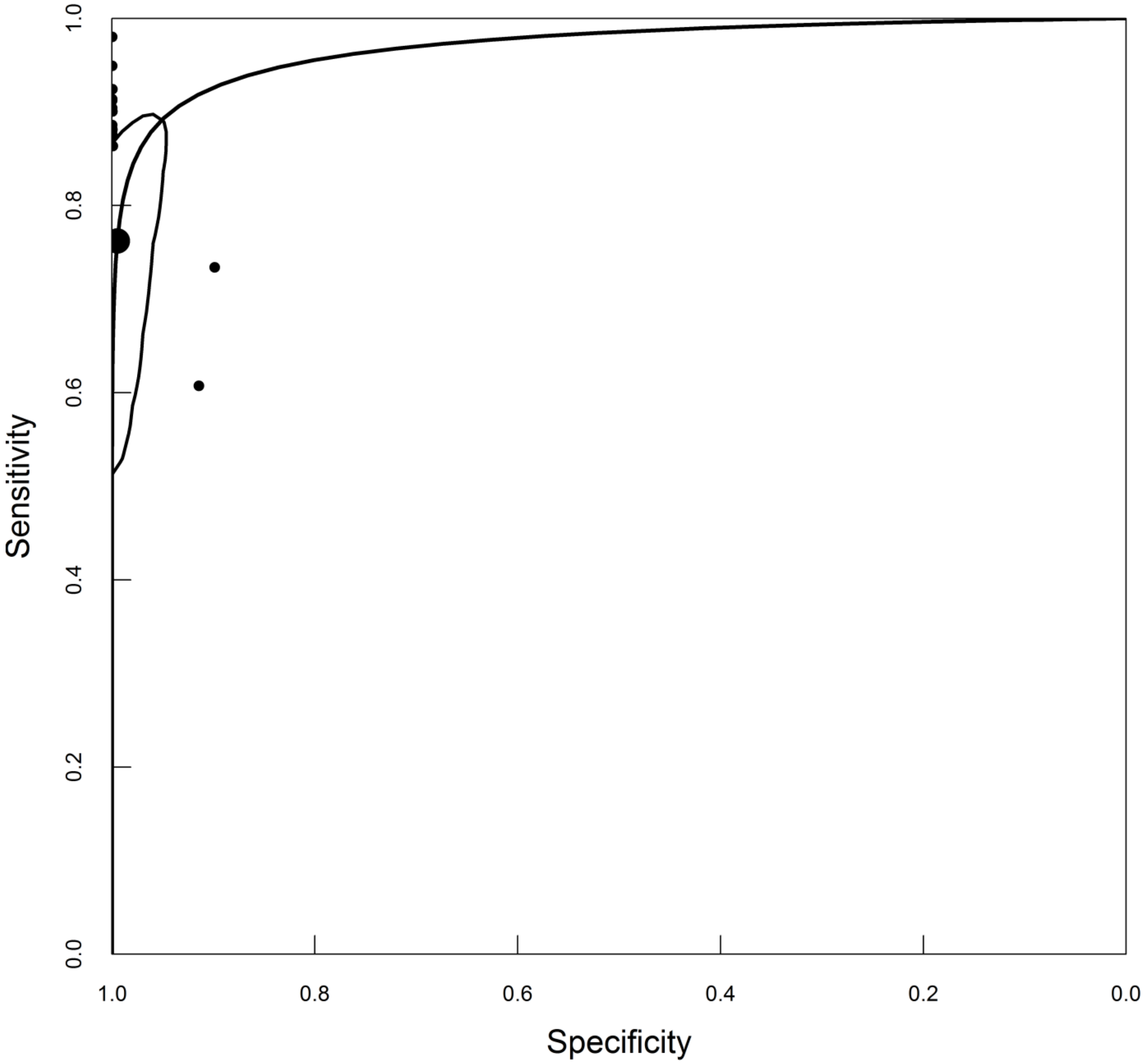
Figure 3. Forest plot of sensitivities and specificities of phenotype versus genotype tests discriminating deficient or intermediate individuals versus others

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ben Salah, L., et al., 2013	8	0	0	80	1.00 [0.63, 1.00]	1.00 [0.95, 1.00]
Fakhoury, M., et al., 2007	12	29	2	77	0.86 [0.57, 0.98]	0.73 [0.63, 0.81]
Fangbin, Z., et al., 2012	5	8	0	287	1.00 [0.48, 1.00]	0.97 [0.95, 0.89]
Ford, L., Graham, V., Berg, J., 2006	29	7	7	359	0.81 [0.64, 0.92]	0.98 [0.96, 0.99]
Ganiere-Monteil, C., et al., 2004	38	3	0	427	1.00 [0.91, 1.00]	0.99 [0.98, 1.00]
Gazouli, M., et al., 2012	24	22	0	62	1.00 [0.86, 1.00]	0.74 [0.63, 0.83]
Hindorf, U., Appell, M.L., 2012	663	291	78	6163	0.89 [0.87, 0.92]	0.95 [0.95, 0.96]
Jorquera, A., et al., 2012	15	3	1	181	0.94 [0.70, 1.00]	0.98 [0.95, 1.00]
Langley, P.G., et al., 2002	6	3	4	40	0.60 [0.26, 0.88]	0.93 [0.81, 0.99]
Larussa, T., et al., 2012	4	14	1	32	0.80 [0.28, 0.99]	0.70 [0.54, 0.82]
Lennard, L., et al., 2013	97	79	9	930	0.92 [0.84, 0.96]	0.92 [0.90, 0.94]
Liang, J. J., et al., 2013	9	6	1	77	0.90 [0.55, 1.00]	0.93 [0.85, 0.97]
Loennechen, T., et al., 1998	23	1	0	236	1.00 [0.85, 1.00]	1.00 [0.98, 1.00]
Ma, X.L., 2006	21	10	1	598	0.95 [0.77, 1.00]	0.98 [0.97, 0.99]
Marinaki, A.M., et al., 2003	5	4	0	76	1.00 [0.48, 1.00]	0.95 [0.88, 0.99]
Milek, M., et al., 2006	6	6	2	81	0.75 [0.35, 0.97]	0.93 [0.86, 0.97]
Schaeffeler, E., et al., 2004	112	17	6	1079	0.95 [0.89, 0.98]	0.98 [0.98, 0.99]
Schwab, M., et al., 2006	5	0	3	85	0.63 [0.24, 0.91]	1.00 [0.96, 1.00]
Serpe, L., et al., 2009	41	265	11	626	0.79 [0.65, 0.89]	0.70 [0.67, 0.73]
Spire-Vayron de la Moureyre, C., et al., 1998a	15	3	1	16	0.94 [0.70, 1.00]	0.84 [0.60, 0.97]
Spire-Vayron de la Moureyre, C., et al., 1998b	18	15	9	149	0.67 [0.46, 0.83]	0.91 [0.85, 0.95]
von Ahsen, N., et al., 2005	5	16	0	50	1.00 [0.48, 1.00]	0.76 [0.64, 0.85]
Wennerstrand, P., et al., 2013	4	19	0	28	1.00 [0.40, 1.00]	0.60 [0.44, 0.74]
Winter, J.W., et al., 2007	11	6	0	113	1.00 [0.72, 1.00]	0.95 [0.89, 0.98]
Xin, H., et al., 2009	7	17	0	126	1.00 [0.59, 1.00]	0.88 [0.82, 0.93]
Yates, C.R., et al., 1997	26	1	0	21	1.00 [0.87, 1.00]	0.95 [0.77, 1.00]
Zhang, L., et al., 2007	7	12	0	259	1.00 [0.59, 1.00]	0.96 [0.92, 0.98]



The forest plots in Figure 3 show 12 out of 27 studies with perfect sensitivity. However, the 95% CIs of these estimates were much smaller than for the sensitivities in Figure 2. This was a result of the higher prevalence of the combined low and intermediate TPMT activity. The specificities were observed to be large, but only two out of 13 studies had perfect specificity. As with Figure 2, the 95% CIs for specificity were narrow. The forest plots in Figure 3 show more heterogeneity than the plots in Figure 2, further justifying the use of a random effects model to account for between-test variability.

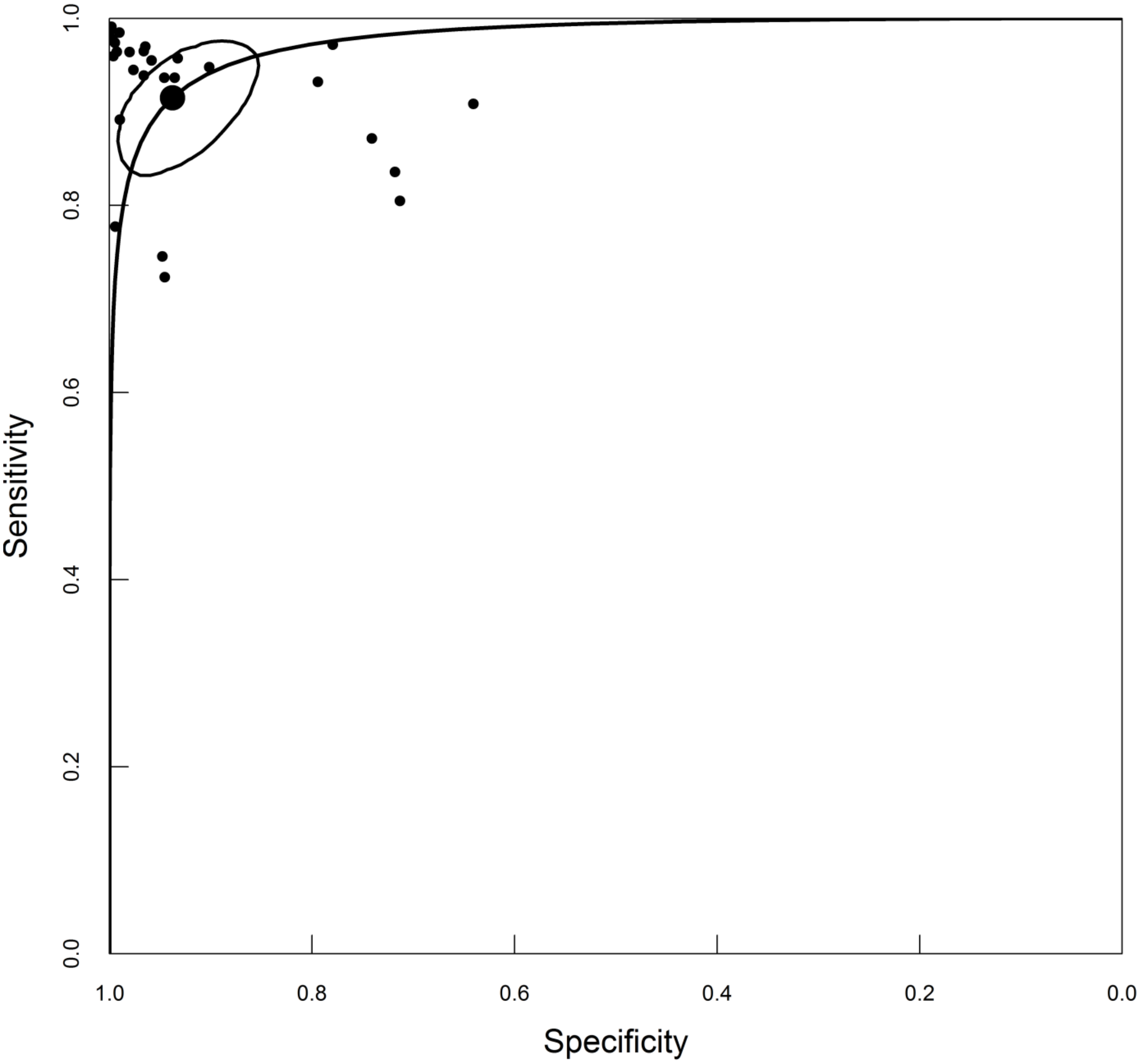
Figure 4. Hierarchical summary ROC curve for the phenotype test discriminating deficient TPMT individuals versus others



The SROC curve was estimated from latent class meta-analysis model assuming imperfect reference standards. Small dots represent the sensitivity and specificity of individual studies and the large dot represents the pooled sensitivity and specificity. The ellipse around the pooled sensitivity and specificity represented the 95% credible region for the pooled sensitivity and specificity.

Based on the latent class model, the pooled sensitivity and specificity of phenotype testing was 75.9% (95% CrI, 58.3% to 87.0%) and 98.9% (95% CrI, 96.3% to 100%), respectively. Figure 4 provides the summary ROC curve from the latent class model, together with the pooled sensitivity and specificity value and associated 95% credible region. The latent class meta-analysis model also provided pooled sensitivity and specificity of the genotype tests. For genotype tests evaluating TPMT*2 and TPMT*3, the pooled sensitivity and specificity was 90.4% (95% CrI, 79.1% to 99.4%) and 100.0% (95% CrI, 99.9% to 100%), respectively. For genotype tests evaluating TPMT*2, TPMT*3 and more polymorphisms, the pooled sensitivity and specificity was 80.7% (95% CrI, 41.7% to 99.4%) and 99.9% (95% CrI, 99.7% to 100%), respectively.

Figure 5. Hierarchical summary ROC curve for the phenotype test discriminating individuals with deficient or intermediate TPMT activity versus others



The SROC curve is estimated from latent class meta-analysis model assuming imperfect reference standards. Small dots represent the sensitivity and specificity of individual studies and the large dot represents the pooled sensitivity and specificity. The ellipse around the pooled sensitivity and specificity represented the 95% credible region for the pooled sensitivity and specificity.

Based on the latent class model, the pooled sensitivity and specificity of phenotype testing was 91.3% (95% CrI, 86.4% to 95.5%) and 92.6% (95% CrI, 86.5% to 96.6%), respectively, when testing individuals with TPMT deficiency or intermediate activity and/or with homozygous or heterozygous TPMT mutations versus the remainder of the population. Figure 5 provides the summary ROC curve from the latent class model, together with the pooled sensitivity and specificity value and associated 95% credible region. The latent class meta-analysis model also provided pooled sensitivity and specificity of the genotype tests. For genotype tests evaluating TPMT*3 mutations only, the pooled sensitivity and specificity was 66.8% (95% CrI, 51.1% to 94.6%) and 99.9% (95% CrI, 99.5% to 100%), respectively. For genotype tests evaluating TPMT*2 and TPMT*3, the pooled sensitivity and specificity was 88.9% (95% CrI, 81.6% to 97.5%) and 99.2% (95% CrI, 98.4% to 99.9%), respectively. For genotype tests evaluating TPMT*2, TPMT*3, and more polymorphisms, the pooled sensitivity and specificity was 93.5% (95% CrI, 84.9% to 99.3%) and 99.9% (95% CrI, 99.7% to 100%), respectively.

4 DISCUSSION

4.1 Diagnostic test accuracy

A total of 30 studies of high quality were identified that compared phenotype testing of TPMT enzyme activity to genotype testing of TPMT mutations, and 27 contained enough information to extract 2x2 contingency tables. Of the 13 studies discriminating individuals with TPMT deficiency versus the remainder of the population, ten studies reported on genotype tests evaluating only TPMT*2 and TPMT*3, and three studies reported on genotype tests evaluating TPMT*2, TPMT*3, and more polymorphisms. Studies discriminating individuals with TPMT deficiency or intermediate activity versus the remainder of the population were grouped into three categories for the phenotype test: two studies reported on genotype tests evaluating only TPMT*3, 19 studies reported on genotype tests evaluating TPMT*2 and TPMT*3, and six studies reported on genotype tests evaluating TPMT*2, TPMT*3, and more polymorphisms.

For the test discriminating between patients with deficient TPMT enzyme activity or a homozygous TPMT mutation and the rest of the population, the individual study estimates of sensitivity had wide 95% CIs. This is a result of the small number of patients with deficient TPMT enzyme activity or a homozygous TPMT mutation included in the studies. In 13 studies with a total of 10,956 patients, only 69 patients had deficient TPMT enzyme activity or a homozygous TPMT mutation. However, as shown in Fig. 2, although the sensitivity estimates were uncertain, the specificity estimates of individual studies demonstrated narrow 95% CIs and the estimates were close to one. The pooled sensitivity was 97.1% [CI: 89.9%, 99.6%], and pooled specificity was 99.8% [CI: 99.7%, 99.9%].

Surprisingly, when distinguishing deficient TPMT activity from the remainder of the population, the genotype tests evaluating TPMT*2, TPMT*3, and more polymorphisms had a *lower* sensitivity than genotype tests evaluating TPMT*2 and TPMT*3 only. When distinguishing deficient or intermediate TPMT activity from the remainder of the population, as expected, the genotype test evaluating TPMT*2 and TPMT*3 had *higher* sensitivity than the test evaluating TMPT*3 only and *lower* sensitivity than the test evaluating TPMT*2, TPMT*3, and more. Because of the large 95% CrI associated with the sensitivity estimates however, none of differences achieved statistical significance. As a result, the meta-analysis was not able to identify improved clinical validity when testing for more polymorphisms than TPMT*2 and TPMT*3 alone.

A previous systematic review by Donnan et al. reported ranges for the sensitivity and specificity of the TPMT genotype test of 55% to 100% and 94% to 100%, respectively. The TPMT phenotype test sensitivity and specificity ranged from 92% to 100% and from 86% to 98%, respectively (6). A meta-analysis by Booth et al. for discriminating between patients with low or intermediate TPMT enzyme activity (homozygous or heterozygous TPMT mutations) and the rest of the population reported a sensitivity range of 70.33% to 86.15% (lower-bound 95% CI, 54.52% to 70.88%; upper-bound CI, 78.50% to 96.33% for phenotype tests and a pooled estimate of 79.90% (95% CI, 74.81% to 84.55%) for genotype tests (64). Due to the rarity of the homozygous mutation, Booth et al. did not meta-analyze tests discriminating between patients with deficient TPMT enzyme activity (homozygous TPMT mutation) and the rest of the population.

Although there are many more studies of testing patients with low or intermediate TPMT enzyme activity (homozygous or heterozygous TPMT mutation) versus the rest of the population, these studies do not address the most pressing clinical issue: identifying patients who are at greatly increased risk of ADEs from thiopurine drugs. Although the test for low or intermediate TPMT enzyme activity or a homozygous or heterozygous TPMT mutation is the most feasible study due to the low prevalence of heterozygous patients, it does not provide guidance on whether doses of thiopurines should be reduced, or whether the drugs should be avoided altogether. Generation of the ROC curve with better estimates of sensitivity and specificity could only be obtained by testing a sample large enough to include more cases with the homozygous mutation, which is often infeasible. This underscores the challenge of determining accurate performance metrics and clinical decision-making cut-points for any rare genetic variants that are implicated in drug metabolism – a challenge that is expected to become more common as the field of pharmacogenomics expands.

4.2 Meta-analysis of diagnostic test results

A latent class meta-analysis statistical method was used to estimate the sensitivity and specificity of phenotype and genotype TPMT tests. The method used was sensitive to starting values, but it was found that results either converged to reasonable results or to boundary conditions that were highly unrealistic (e.g., sensitivity = 0, specificity = 1). Another method was available with different model assumptions (the conditionally dependent model) but the software would not find a solution that fit the data (i.e., could not converge).

It is clear that meta-analysis of diagnostic test accuracy should not consist of simply pooling or averaging sensitivity and specificity values. This study has shown that statistical methods are available that can address the heterogeneity of DTA studies, can address the correlation of sensitivity and specificity estimates, and can also address the issue of imperfect references standards should they exist. Although computationally intensive, these analyses were not technically challenging to implement (63, 65).

4.3 Strengths and limitations

A strength of this study was that the estimates of phenotype and genotype sensitivity and specificity were calculated simultaneously. The latent class meta-analysis method used addresses the correlation of sensitivity and specificity values as well as the imperfect nature of the available reference standards. A full SROC curve was estimated for phenotype TPMT testing as well. Another strength was that due to the relatively large number of available studies it was possible to restrict the meta-analysis to 30 high-quality studies.

A limitation of the study was that the judgment of high-quality was determined by the reviewers. The QUADAS-2 tool used to appraise the literature was not designed to classify studies as high or low quality, or to assign a numeric score. Rather, the tool allowed the reviewers to summarize issues of bias and applicability. Other reviewers might have established different definitions of high quality and obtained different results. However, strengths of the QUADAS-2 were that is is recommended by the Cochrane group (66), it is open-ended and allowed for inclusion of a customized genomics domain, and the tool allows for a systematic evaluation of the risk of bias and lack of applicability of individual studies. Another study limitation was that there were not enough studies to conduct stratified meta-analysis by ethnic group. It is known that different ethnicities have different proportions of TPMT polymorphisms and this may result in different values of sensitivity and specificity for the genotype test. It may be inappropriate to apply accuracy results derived from pooling studies of heterogeneous populations to clinical decision -making for specific ethnic groups. Finally, the meta-analysis was forced to assume conditional independence between the phenotype and genotype tests. The conditionally dependent model is likely more appropriate because phenotype and genotype test results are expected to be correlated since the genotype test measures polymorphisms of the gene that codes for the

TPMT enzyme whose activity is measured by the phenotype test. Due to challenges with results that would not converge, the conditionally dependent model could not be used.

4.4 Implications for clinical practice and research

The pooled estimates of sensitivity suggest that genotype testing has higher sensitivity than phenotype testing as long as both TPMT*2 and TPMT*3 polymorphisms are tested. A high value for sensitivity is important in diagnostic applications to rule out the presence of deficiencies that may be associated with drug-related toxicity, thus allowing full therapeutic doses of thiopurines to be administered. However, due to the large 95% CrIs surrounding the sensitivity values, the pooled estimate remains uncertain. Both the genotype and phenotype tests demonstrated high specificity. A high specificity is useful for ruling in the presence of a deficiency.

The meta-analysis reflected the wide uncertainty that is evident in the TPMT diagnostic performance literature and demonstrated that one cannot conclude that one test is superior to the other. The question of diagnostic test accuracy is best addressed in large population studies of targeted ethnic groups, with sufficient numbers of patients with homozygous mutations to enable the determination of stable estimates.

Despite the uncertainty surrounding the test accuracy results, the pooled estimates for sensitivity and specificity of TPMT phenotype and genotype testing, together with 95% CrI information, are valuable for economic evaluations comparing alternative testing approaches and testing technologies. A probabilistic sensitivity analysis that simultaneously incorporates the pooled estimates and credible intervals could be undertaken to provide an indication of the relative cost-effectiveness of one approach over another. Value of information methods may also be undertaken to determine the cost associated with collecting additional data to reduce the observed uncertainty (67).

The field of pharmacogenetic testing continues to grow and evolve, allowing more patients to benefit from a personalized approach to drug selection and dosing (12). Deoxyribonucleic acid (DNA)-testing technologies are also evolving rapidly with the advent of next generation sequencing that allows for simultaneous disease diagnostic testing, pharmacogenetic testing as well as screening for the risk of future diseases (65). The positioning of pharmacogenetic testing for TPMT and other enzymatic deficiencies in the larger context of next generation sequencing is an area for future research.

5 CONCLUSION

When TPMT testing was used to identify individuals with deficient TPMT enzyme activity (homozygous TPMT mutation), the sensitivity and specificity of the phenotype test was 75.9% and 98.9% with CrI of 58.3% to 87.0% and 96.3% to 100%, respectively. The sensitivity and specificity of the genotype test with TPMT*2 and TPMT*3 polymorphisms was 90.4% and 100% with CrI of 79.1% to 99.4% and 99.9% to 100%, respectively. The sensitivity and specificity of the genotype test with more polymorphisms was 80.7% and 99.9% with CrI of 41.7% to 99.4% and 99.7% to 100%, respectively.

When TPMT testing was used to identify individuals with deficient to intermediate TPMT enzyme activity (homozygous or heterozygous TPMT mutations), the sensitivity and specificity of the phenotype test was 91.3% and 92.6% with CrI of 86.4% to 95.5% and 86.5% to 96.6%, respectively. The sensitivity and specificity of the genotype test with TPMT*2 and TPMT*3 polymorphisms was 88.9% and 99.2% with CrI of 81.6% to 97.5% and 98.4% to 99.9%, respectively. The sensitivity and specificity of the genotype test with more polymorphisms was 93.5% and 99.9% with CrI of 84.9% to 99.3% and 99.7% to 100%, respectively. The sensitivity and specificity of the genotype test with only the TPMT*3 polymorphism tested was 66.8% and 99.9% with CrI of 51.1% to 94.6% and 99.5% to 100%.

The pooled estimates of sensitivity suggest that genotype testing has higher sensitivity than phenotype testing as long as both TPMT*2 and TPMT*3 polymorphisms are tested. However, due to the large 95% CrIs around sensitivity estimates the results remain uncertain. Both tests have been shown to have high specificity. Therefore, this meta-analysis cannot conclude that one test is superior to the other. The methods applied in this research were considered the most appropriate for the problem. Although the methods were statistically complex, software was available that made the implementation straight-forward. The sensitivity and specificity of the phenotype and genotype tests determined using these methods were quite different from a simple pooling of sensitivity and specificity. Therefore, HSROC or bivariate methods are recommended for DTA meta-analyses rather simple pooling.

REFERENCES

1. Roy LM, Ungar WJ, Zur RM. Thiopurine S-methyltransferase testing for averting drug toxicity in patients receiving thiopurines: A systematic review and quality appraisal. 2015 March 26, 2015. Report No.: 2015-02.
2. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. Canadian Medical Association Journal. 2004;170(11):1678-86.
3. Stanulla M, Schaeffeler E, Flohr T, Cario G, Schrauder A, Zimmermann M, et al. Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia. JAMA. 2005;293(12):1485-9. Epub 2005/03/24.
4. Alves S, Amorim A, Ferreira F, Prata MJ. Influence of the variable number of tandem repeats located in the promoter region of the thiopurine methyltransferase gene on enzymatic activity. Clinical Pharmacology & Therapeutics. 2001;70(2):165-74.
5. Anglicheau D, Sanquer S, Loriot MA, Beaune P, Thervet E. Thiopurine methyltransferase activity: new conditions for reversed-phase high-performance liquid chromatographic assay without extraction and genotypic-phenotypic correlation. Journal of Chromatography B: Analytical Technologies in the Biomedical & Life Sciences. 2002;773(2):119-27.
6. Donnan JR, Ungar WJ, Mathews M, Rahman P. Systematic review of thiopurine methyltransferase genotype and enzymatic testing strategies. Ther Drug Monit. 2011;33(2):192-9.
7. Indjova D, Shipkova M, Atanasova S, Niedmann PD, Armstrong VW, Svinarov D, et al. Determination of thiopurine methyltransferase phenotype in isolated human erythrocytes using a new simple nonradioactive HPLC method. Therapeutic Drug Monitoring. 2003;25(5):637-44.
8. Kham SKY, Soh CK, Liu TC, Chan YH, Ariffin H, Tan PL, et al. Thiopurine S-methyltransferase activity in three major Asian populations: a population-based study in Singapore. Eur J Clin Pharmacol. 2008;64(4):373-9.
9. Larovere LE, De Kremer RD, Lambooy LHJ, De Abreu RA. Genetic polymorphism of thiopurine S-methyltransferase in Argentina. Ann Clin Biochem. 2003;40(4):388-93.
10. Winter JW, Gaffney D, Shapiro D, Spooner RJ, Marinaki AM, Sanderson JD, et al. Assessment of thiopurine methyltransferase enzyme activity is superior to genotype in predicting myelosuppression following azathioprine therapy in patients with inflammatory bowel disease. Alimentary Pharmacology & Therapeutics. 2007;25(9):1069-77.
11. Wusk B, Kullak-Ublick GA, Rammert C, von Eckardstein A, Fried M, Rentsch KM. Thiopurine S-methyltransferase polymorphisms: efficient screening method for patients considering taking thiopurine drugs. Eur J Clin Pharmacol. 2004;60(1):5-10.
12. Heckmann JM, Lambson EMT, Little F, Owen EP. Thiopurine methyltransferase (TPMT) heterozygosity and enzyme activity as predictive tests for the development of azathioprine-related adverse events. Journal of the neurological sciences. 2005;231(1-2):71-80.
13. Ujiie S, Sasaki T, Mizugaki M, Ishikawa M, Hiratsuka M. Functional characterization of 23 allelic variants of thiopurine S-methyltransferase gene (TPMT*2 - *24). Pharmacogenetics and genomics. 2008;18(10):887-93.
14. Jones CD, Smart C, Titus A, Blyden G, Dorvil M, Nwadike N. Thiopurine methyltransferase activity in a sample population of black subjects in Florida. Clin Pharmacol Ther. 1993;53(3):348-53.
15. Lee EJ, Kalow W. Thiopurine S-methyltransferase activity in a Chinese population. Clin Pharmacol Ther. 1993;54(1):28-33.
16. Park-Hah JO, Klemetsdal B, Lysaa R, Choi KH, Aarbakke J. Thiopurine methyltransferase activity in a Korean population sample of children. Clin Pharmacol Ther. 1996;60(1):68-74.
17. Booth RA, Ansari MT, Tricco AC, Loit E, Weeks L, Doucette S, et al. Assessment of thiopurine methyltransferase activity in patients prescribed azathioprine or other thiopurine-based drugs. Evidence Report/Technology Assessment No. 196. Prepared by the University of Ottawa Evidence-based Practice Center under Contract No. 290-2007-10059-I AHRQ Publication No. 11-E002. Rockville, MD: Agency for Healthcare Research and Quality.: 2010.
18. Donnan JR, Ungar WJ, Mathews M, Hancock-Howard RL. Health Technology Assessment of Thiopurine Methyltransferase Testing for Guiding 6-Mercaptopurine Doses in Pediatric Patients with Acute Lymphoblastic Leukemia. Technology Assessment at SickKids (TASK), 2010.

19. Donnan JR, Ungar WJ, Mathews M, Hancock-Howard RL, Rahman P. A cost effectiveness analysis of thiopurine methyltransferase testing for guiding 6-mercaptopurine dosing in children with acute lymphoblastic leukemia. Pediatr Blood Cancer. 2011;57(2):231-9.
20. Centre for Reviews and Dissemination. Systematic reviews of clinical tests. In: York Uo, editor. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: York Publishing Services; 2009.
21. Deeks JJ. Using evaluations of diagnostic tests: understanding their limitations and making the most of available evidence. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 1999;10(7):761-8.
22. Hurley J. Meta-analysis of clinical studies of diagnostic tests: developments in how the receiver operating characteristic "works". Archives of pathology & laboratory medicine. 2011;135(12):1585-90.
23. Group DTAW. Handbook for DTA Reviews. The Cochrane Collaboration; 2013 [cited 2013 March 25]; Available from: http://srdta.cochrane.org/handbook-dta-reviews.
24. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003;59(4):936-46.
25. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR American journal of roentgenology. 2006;187(2):271-81.
26. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in medicine. 2001;20(19):2865-84.
27. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001;57(1):158-67.
28. Albert PS. Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. Statistics in medicine. 2009;28(5):780-97.
29. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. Biometrics. 2012;68(4):1285-93.
30. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health technology assessment. 2007;11(50):iii, ix-51.
31. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology. 2005;58(10):982-90.
32. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Medical decision making : an international journal of the Society for Medical Decision Making. 2008;28(5):621-38.
33. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.
34. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
35. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. Statistics in medicine. 2009;28(3):441-61.
36. Sinclair A, Xie X, Teltscher M, Dendukuri N. Systematic review and meta-analysis of a urine-based pneumococcal antigen test for diagnosis of community-acquired pneumonia caused by Streptococcus pneumoniae. Journal of clinical microbiology. 2013;51(7):2303-10.
37. Ben Salah L, Belkhiria el Haj Amor M, Chbili C, Khlifi S, Fathallah N, Bougmiza I, et al. Analysis of thiopurine S-methyltransferase phenotype-genotype in a Tunisian population with Crohn's disease. Eur J Drug Metab Pharmacokinet. 2013;38(4):241-4.
38. Fakhoury M, Andreu-Gallien J, Mahr A, Medard Y, Azougagh S, Vilmer E, et al. Should TPMT genotype and activity be used to monitor 6-mercaptopurine treatment in children with acute lymphoblastic leukaemia? J Clin Pharm Ther. 2007;32(6):633-9.
39. Ford L, Graham V, Berg J. Whole-blood thiopurine S-methyltransferase activity with genotype concordance: a new, simplified phenotyping assay. Ann Clin Biochem. 2006;43(Pt 5):354-60.
40. Ganiere-Monteil C, Medard Y, Lejus C, Bruneau B, Pineau A, Fenneteau O, et al. Phenotype and genotype for thiopurine methyltransferase activity in the French Caucasian population: impact of age. Eur J Clin Pharmacol. 2004;60(2):89-96.

41. Gazouli M, Pachoula I, Panayotou I, Chouliaras G, Anagnou NP, Chroussos G, et al. Thiopurine methyltransferase genotype and thiopurine S-methyltransferase activity in Greek children with inflammatory bowel disease. Annals of Gastroenterology. 2012;25(3):249-53.
42. Larussa T, Suraci E, Lentini M, Nazionale I, Gallo L, Abenavoli L, et al. High prevalence of polymorphism and low activity of thiopurine methyltransferase in patients with inflammatory bowel disease. Eur. 2012;23(3):273-7.
43. Schaeffeler E, Fischer C, Brockmeier D, Wernet D, Moerike K, Eichelbaum M, et al. Comprehensive analysis of thiopurine S-methyltransferase phenotype-genotype correlation in a large population of German-Caucasians and identification of novel TPMT variants. Pharmacogenetics. 2004;14(7):407-17.
44. Schwab M, Schaffeler E, Marx C, Fischer C, Lang T, Behrens C, et al. Azathioprine therapy and adverse drug reactions in patients with inflammatory bowel disease: impact of thiopurine S-methyltransferase polymorphism. Pharmacogenetics. 2002;12(6):429-36.
45. Serpe L, Calvo PL, Muntoni E, D'Antico S, Giaccone M, Avagnina A, et al. Thiopurine S-methyltransferase pharmacogenetics in a large-scale healthy Italian-Caucasian population: differences in enzyme activity. Pharmacogenomics. 2009;10(11):1753-65.
46. Yates CR, Krynetski EY, Loennechen T, Fessing MY, Tai HL, Pui CH, et al. Molecular diagnosis of thiopurine S-methyltransferase deficiency: genetic basis for azathioprine and mercaptopurine intolerance. Annals of Internal Medicine. 1997;126(8):608-14.
47. Hindorf U, Appell ML. Genotyping should be considered the primary choice for pre-treatment evaluation of thiopurine methyltransferase function. J Crohns Colitis. 2012;6(6):655-9.
48. Spire-Vayron de la Moureyre C, Debuysere H, Mastain B, Vinner E, Marez D, Lo Guidice JM, et al. Genotypic and phenotypic analysis of the polymorphic thiopurine S-methyltransferase gene (TPMT) in a European population. Br J Pharmacol. 1998;125(4):879-87.
49. Spire-Vayron de la Moureyre C, Debuysere H, Sabbagh N, Marez D, Vinner E, Chevalier ED, et al. Detection of known and new mutations in the thiopurine S-methyltransferase gene by single-strand conformation polymorphism analysis. Human mutation. 1998;12(3):177-85.
50. Langley PG, Underhill J, Tredger JM, Norris S, McFarlane IG. Thiopurine methyltransferase phenotype and genotype in relation to azathioprine therapy in autoimmune hepatitis. J Hepatol. 2002;37(4):441-7.
51. Lennard L, Cartwright CS, Wade R, Richards SM, Vora A. Thiopurine methyltransferase genotype-phenotype discordance and thiopurine active metabolite formation in childhood acute lymphoblastic leukaemia. British Journal of Clinical Pharmacology. 2013;76(1):125-36.
52. Fangbin Z, Xiang G, Minhu C, Liang D, Feng X, Min H, et al. Should Thiopurine Methyltransferase Genotypes and Phenotypes be Measured Before Thiopurine Therapy in Patients With Inflammatory Bowel Disease? Therapeutic drug monitoring. 2012;34(6):695-701 10.1097/FTD.0b013e3182731925.
53. Liang JJ, Geske JR, Boilson BA, Frantz RP, Edwards BS, Kushwaha SS, et al. TPMT genetic variants are associated with increased rejection with azathioprine use in heart transplantation. Pharmacogenetics and genomics. 2013;23(12):658-65.
54. Ma XL, Wu MY, Hu YM, Zu P, Li ZG. Relationships between thiopurine methyltransferase gene polymorphisms and its enzymatic activity. [Chinese]. Zhonghua zhong liu za zhi [Chinese journal of oncology]. 2006;28(6):456-9.
55. Marinaki AM, Arenas M, Khan ZH, Lewis CM, Shobowale-Bakre E-M, Escuredo E, et al. Genetic determinants of the thiopurine methyltransferase intermediate activity phenotype in British Asians and Caucasians. Pharmacogenetics. 2003;13(2):97-105.
56. Milek M, Murn J, Jaksic Z, Lukac Bajalo J, Jazbec J, Mlinaric Rascan I. Thiopurine S-methyltransferase pharmacogenetics: genotype to phenotype correlation in the Slovenian population. Pharmacology. 2006;77(3):105-14.
57. von Ahsen N, Armstrong VW, Behrens C, von Tirpitz C, Stallmach A, Herfarth H, et al. Association of inosine triphosphatase 94C>A and thiopurine S-methyltransferase deficiency with adverse events and study drop-outs under azathioprine therapy in a prospective Crohn disease study.[Erratum appears in Clin Chem. 2006 Aug;52(8):1628 Note: Schutz, Ekkehard [added]]. Clinical chemistry. 2005;51(12):2282-8.

58. Wennerstrand P, Martensson LG, Soderhall S, Zimdahl A, Appell ML. Methotrexate binds to recombinant thiopurine S-methyltransferase and inhibits enzyme activity after high-dose infusions in childhood leukaemia. Eur J Clin Pharmacol. 2013;69(9):1641-9.
59. Xin H-W, Xiong H, Wu X-C, Li Q, Xiong L, Yu A-R. Relationships between thiopurine S-methyltransferase polymorphism and azathioprine-related adverse drug reactions in Chinese renal transplant recipients. Eur J Clin Pharmacol. 2009;65(3):249-55.
60. Zhang L-R, Song D-K, Zhang W, Zhao J, Jia L-J, Xing D-L. Efficient screening method of the thiopurine methyltransferase polymorphisms for patients considering taking thiopurine drugs in a Chinese Han population in Henan Province (central China). Clinica Chimica Acta. 2007;376(1-2):45-51.
61. Jorquera A, Solari S, Vollrath V, Guerra I, Chianale J, Cofre C, et al. [Phenotype and genotype of thiopurine methyltransferase in Chilean individuals]. Rev Med Chil. 2012;140(7):889-95. Genotipo y fenotipo de la enzima tiopurina metiltransferasa en poblacion chilena.
62. Loennechen T, Yates CR, Fessing MY, Relling MV, Krynetski EY, Evans WE. Isolation of a human thiopurine S-methyltransferase (TPMT) complementary DNA with a single nucleotide transition A719G (TPMT*3C) and its association with loss of TPMT protein and catalytic activity in humans. Clinical Pharmacology & Therapeutics. 1998;64(1):46-51.
63. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing. 2000;10(4):325-37.
64. Booth RA, Ansari MT, Loit E, Tricco AC, Weeks L, Doucette S, et al. Assessment of thiopurine S-methyltransferase activity in patients prescribed thiopurines: a systematic review. Annals of Internal Medicine. 2011;154(12):814-23, W-295-8.
65. Dendukuri N. Software. 2015 [cited 2015 June 23]; Available from: http://www.nandinidendukuri.com/index.php?option=com_content&view=category&id=41&Itemid=60.
66. Briggs AH, Claxton K, Sculpher MJ. Decision Modelling for Health Economic Evaluation: Oxford University Press; 2006.
67. Reitsma J, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. 2009 October 27, 2009. Report No.