The Hospital for Sick Children
Technology Assessment at SickKids (TASK)

FULL REPORT

THIOPURINE S-METHYLTRANSFERASE TESTING FOR AVERTING DRUG TOXICITY IN PATIENTS RECEIVING THIOPURINES: A SYSTEMATIC REVIEW AND QUALITY APPRAISAL

Authors:

Lilla M. Roy, RN, BScN, MSc
Clinical Research Project Coordinator, Child Health Evaluative Services, The Hospital for Sick Children Peter Gilgan Centre for Research and Learning, Toronto

Wendy J. Ungar, MSc, PhD
Senior Scientist, Child Health Evaluative Sciences, The Hospital for Sick Children Peter Gilgan Centre for Research and Learning, Toronto; Professor, Health Policy, Management & Evaluation, University of Toronto

Richard M. Zur, PhD
Research Project Manager, Child Health Evaluative Services, The Hospital for Sick Children Peter Gilgan Centre for Research and Learning, Toronto

Corresponding Author:

Wendy J. Ungar, MSc, PhD
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

Report No. 2015-02

Date: July 29, 2015

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

Co-investigators:

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

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

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

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 the following individuals who provided assistance with study translation: Fernando Ortiz, Ella Hyatt, Zhaoxin Dong, Roland Arnold, Takashi Hirama, Cindy Zhang, Jessie Guo, Simon Tian, Brian Kim, Jennifer Park, and Vera Nenadovic. We would also like to thank Tara Paton and Kassa Beimnet for their technical expertise regarding genotype and phenotype testing, and Cathy Pajunen for her library services and referencing technical support.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

6-MMP	6-methyl-mercaptopurine
6-MP	6-mercaptopurine
6-TG	6-thioguanine
ADE	Adverse drug event
AHRQ	Agency for Healthcare Research and Quality
AiH	Autoimmune hepatitis
ALL	Acute lymphocytic leukemia
APEX	Arrayed primer extension technology
ARMS	Multiplex amplification refractory mutation
AS-PCR	Allele-specific polymerase chain reaction
AZA	Azathioprine
CE	Capillary electrophoresis
CEA	Cost-effectiveness analysis
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CCTR	Cochrane Central Register of Controlled Trials
CDSR	Cochrane Database of Systematic Reviews
DARE	Database of Abstracts of Reviews of Effects
DHPLC	Denaturing high performance liquid chromatography
gHb	Gram of hemoglobin
h	Hour
Hb	Hemoglobin
HPLC	High performance liquid chromatography
HTA	Health Technology Assessment
IBD	Inflammatory Bowel Disease
IPA	International Pharmaceutical Abstracts
ITPA	Inosine triphosphatase
MA	meta-analysis
MALDI-TOF MS	Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass spectrometry
MeSH	Medical Subject Headings
Mg	Milligram
Min	Minute

Ml	Millilitre
MS Access	Microsoft Access
MTG	Methylthioguanine
NHSEED	National Health Service Economic Evaluation Database
Ng	Nanogram
Nmol	Nanomole
NPV	Negative predictive value
NR	Not reported
PCR	Polymerase chain reaction
Pmol	Picomole
pRBC	Packed red blood cells
PPV	Positive predictive value
QA	Quality assurance
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
SSCP	Single strand conformation polymorphism
SNP	Single nucleotide polymorphism
SR	Systematic review
TASK	Technology Assessment at Sickkids
TPMT	Thiopurine s-methyltransferase
U	Unit
UC	Ulcerative colitis
UK	United Kingdom
UV	Ultraviolet

EXECUTIVE SUMMARY

Introduction

Thiopurine S-methyltransferase (TPMT) is an enzyme that metabolizes thiopurine drugs which are commonly used in maintenance treatment for childhood leukemias, as well as, less commonly, for inflammatory bowel disease (IBD), transplant recipients, and dermatological conditions. The absence or a deficiency of TPMT can significantly increase the risk of adverse drug event (ADE) in persons receiving thiopurine therapy as they are unable to metabolize the drug. There has long been phenotype blood testing to measure TPMT enzyme activity, and more recently a genotype test is sued to identify individual as with the genetic variants that determine TPMT activity. Uncertainty remains however, regarding which is the optimal test.

Objectives

The objectives of this study were to systematically review the literature on the performance characteristics of thiopurine testing for TPMT deficiency, to appraise the quality of the literature, and to identify the characteristics of high quality studies.

Methods

A systematic search of electronic databases was conducted, including 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 they must have reported (or provided data to calculate) sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), or concordance between the two technologies.

The abstracts and full text of papers were reviewed to identify studies that met the inclusion criteria. The quality appraisal was completed using the Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS-2). Data extraction from the resulting studies included basic study design characteristics, study results, diagnostic test performance characteristics, and raw data to

populate 2x2 and 3x3 contingency tables to enable the calculation of sensitivity, specificity, PPV, NPV and concordance.

Results

Four thousand seventy-one studies were identified through the database and grey literature search. Three hundred and seventy three records required full text review, and 121 records were reviewed for relevant data. Sixty six studies had sufficient data for inclusion, and underwent quality appraisal. These 66 studies comprised three categories – a category of phenotype-genotype comparisons, and a category of phenotype-phenotype comparisons and genotype-genotype comparisons. In total, 30/55 phenotype-genotype comparisons were designated high quality by the quality appraisal, and 6/11 phenotype-phenotype or genotype-genotype comparisons were designated as high quality.

Studies considered of low quality generally contained unclear information relating to the quality components of the appraisal, as opposed to obvious bias or concerns for applicability. Thirteen of 30 high quality studies had low bias and low concern for applicability, while the remaining high quality studies had at least one domain with unclear or high risk associated with it. All of the high quality studies were published between 1997 and 2013, and examined a range of genotype and phenotype test methods.

Based on available data from 15 studies, the calculated sensitivity for genotyping to identify a homozygous mutation ranged from 0.0% to 100.0% and with data that were available from 26 studies specificity ranged from 97.8% to 100.0%. Based on available data from 25 studies, the calculated sensitivity to detect a homozygous or heterozygous mutation ranged from 13.4 to 100.0% and specificity ranged from 90.9 to 100.0% using data available from 26 studies.

Discussion

The choice of technologies available for the diagnosis of TPMT deficiency is varied. This review revealed a diverse and large body of literature assessing both phenotype and genotype technologies for TPMT testing across several disease states. There are limitations to both genotype testing and phenotype testing, and neither test can be referred to as the ‘gold standard’ for identifying TPMT deficiency.

The quality appraisal revealed that inadequate reporting of count data, descriptive information of index tests, reference tests, and recruitment methods, and study populations largely contributed to the exclusion of studies due to quality. Lack of reporting of diagnostic test accuracy indicates a need for guidance on reporting of test performance characteristics for diagnostic technologies. Thirty high quality studies comparing phenotype and genotype technologies were included in this review. The number of polymorphisms included in genotype tests ranged from two to nine, with most studies including TPMT*2 and TPMT*3, the most common genetic variants in persons with deficient TPMT activity. Among the fifteen studies for which both sensitivity and specificity of genotyping could be calculated, ten demonstrated perfect (100%) sensitivity and specificity. The inference of perfect values is misleading, however. The low prevalence of homozygous mutations (0.3%) made it difficult to generate sample sizes that were large enough for a stable rate of detection of homozygous mutations. The variation in sensitivity and specificity observed in the present review may also be related to the disease context. The tolerance for the risk of serious ADEs, and consequently values for sensitivity and specificity, may be preferred for chronic disease such as IBD and dermatological conditions versus life-threatening disease such as ALL.

Conclusion

There is a growing use of personalized medicine applications such as pharmacogenomics in clinical diagnostics and clinical decision-making for selection of drug treatment and dose. This review of the literature comparing phenotype testing and genotype testing for TPMT status demonstrated a broad base of evidence these tests. The quality of the studies for assessing diagnostic test accuracy was mixed. The low prevalence of patients with deficient TPMT activity or homogeneous TPMT mutations made estimates of sensitivity of the tests uncertain. The accuracy of genotyping is also affected by the range of polymorphisms included in the test. Routine testing for all possible polymorphisms is more costly and unlikely to be feasible for health care institutions. Nevertheless, clinical and institutional decision-makers require high quality evidence of clinical validity and clinical utility of TPMT genotyping technologies to ensure appropriate and consistent use in patient populations who would benefit from this testing.

1 INTRODUCTION

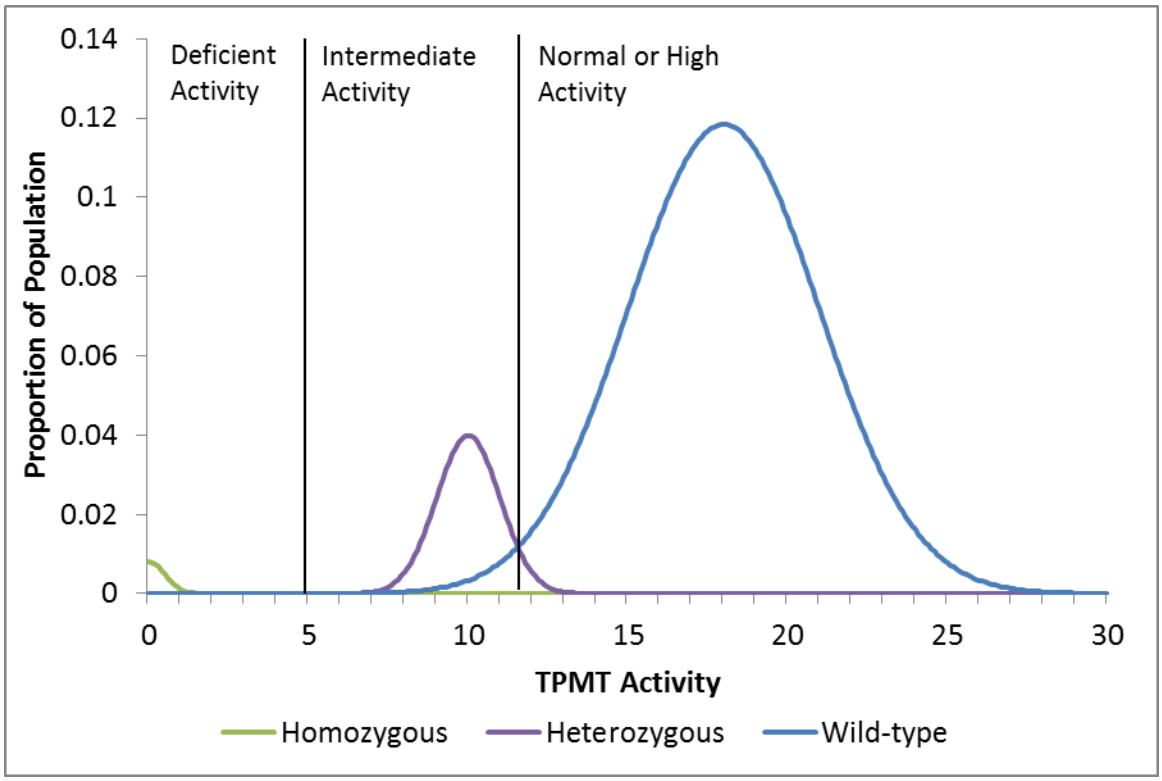
With advances in the field of pharmacogenomics, it is increasingly common to use genetic or biomarker testing to predict an individual's drug responses [1]. This personalized medicine approach allows for more accurate selection of treatments as well as dosing of prescription medicines and the avoidance of potentially serious life-threatening adverse drug events (ADEs). The technologies that are used to test for drug metabolizing enzyme activity and for the presence of genetic variants that affect drug metabolism are rapidly evolving with regard to technical methods as well as scope [2]. This introduces uncertainty for clinical practitioners regarding which tests to use for their patients, and for provincial decision-makers regarding the value for money of these new technologies.

A common application of personalized medicine is testing for deficiency in thiopurine s-methyltransferase (TPMT), the enzyme that metabolizes thiopurines [3]. Thiopurines consist of immunosuppressive and chemotherapeutic drugs that are widely used to treat adults and children with serious conditions, including acute lymphoblastic leukemia (ALL), inflammatory bowel disease (IBD), idiopathic arthritis, and those receiving organ transplants [4]. The clinical consequences of reduced or undetectable TPMT activity are significant. Unless thiopurine drug 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 [5]. These ADEs can result in lengthy hospital admissions and substantial morbidity and reduced quality of life for patients already coping with a serious illness [6, 7]. Approximately 89% of Caucasians have normal (wild type) TPMT activity, 11% are heterozygous with reduced activity, and 0.3% are homozygous with TPMT mutations resulting in undetectable enzyme activity [8, 9]. It is therefore important to identify the presence of TPMT deficiencies in patients prescribed thiopurine drugs.

The incidence of deficient or low activity is rare, while the incidence of 'intermediate' activity and 'high/normal' activity is more common. In contrast to the 'intermediate' and 'high/normal' categories, the 'low' category is phenotypically quite separated from the rest. It is the 'low' category that represents the category of rare variants with significant deficiency that are most at risk for ADE related to thiopurine exposure, and therefore, the most clinically important group to diagnose. The intermediate group is secondarily important, as appropriate dosage in this

population would optimize clinical outcomes, however the determination of the optimal cutpoint differentiating those with 'intermediate' activity from those with 'high' activity is less clear.

Figure 1. Distribution of TPMT activity



There are two approaches to testing for TPMT deficiency. Phenotype tests that measure levels of TPMT enzyme activity *in vitro* are common, but these test results can be confounded by concomitant medications or blood transfusions [2, 10-16]. A genotype test is available that detects the presence of variants in the genes responsible for expressing the TPMT enzyme [17]. While there are 24 genes implicated in TPMT, 3 variants (*3A, *14A and *22) account for 90% of the deficiencies occurring in the population and are the ones commonly included in genetic tests [18]. Patients with these three variants have no detectable enzyme activity. Patients with other variants have approximately 50% of functional enzyme activity [18]. Genetic tests that are designed to detect only the most common variants leave patients with rare mutations at risk [18]. The prevalence of mutations is known to vary by ethnic background [19-22], thus certain segments of the population may be more at risk. It remains uncertain whether an enzyme activity (phenotype) or genotype diagnostic test is the most appropriate strategy for clinical practice.

This uncertainty is especially true in the pediatric population. In children, thiopurine doses are calculated based on weight and serious ADEs may result in significant morbidity [23]. Currently, TPMT deficiency testing varies across Canadian pediatric treatment centres [24]. A survey of Canadian rheumatologists published in 2013 found that 55% of clinicians routinely tested for TPMT deficiency prior to treatment compared to 45% who never tested [24]. Half of the physicians who tested reported avoiding the use of azathioprine (AZA) in patients with deficiency as well as in those patients with reduced activity. Physician uncertainty may be due in part to the disagreement evident in clinical practice guidelines for TPMT testing aimed at different sub-specialties. A recent systematic review (SR) of clinical sub-specialty documents providing guidance on TPMT testing revealed differences in TPMT testing recommendations with five sub-specialty organizations recommending genotyping while four recommended phenotyping. That review also identified differences in thiopurine dosing recommendations when treating patients with identified TPMT deficiencies [25]. In summary, knowledge gaps persist regarding TPMT deficiency testing. Improving information regarding the clinical validity and performance characteristics of TPMT testing strategies will facilitate decision making in the optimal use of TPMT testing for diagnosis and treatment with thiopurines.

1.1 Objectives

The aim of this study was to evaluate the evidence regarding TPMT diagnostic testing. The research objectives were:

1. To systematically review the literature on the performance characteristics of phenotype testing and genotype testing for TPMT deficiency.

2. To appraise the quality of the TPMT testing literature and to identify the characteristics of high quality studies.

2 METHODS

This section describes the methods used for the literature review of the performance characteristics of phenotype testing and genotype testing for TPMT deficiency. The inclusion and exclusion criteria, the literature search design, and data extraction strategy are described. Subsequently, the quality appraisal methods are described, including decision rules, assumptions, and the addition of a genomics domain.

2.1 Systematic review

2.1.1 Inclusion and exclusion criteria

Inclusion criteria specified any study conducted in humans that evaluated either a TPMT genotype or TPMT phenotype technology in comparison to a reference standard, where reference standard was indicated as another phenotype or genotype test such that the comparison could be phenotype-phenotype, genotype-genotype, or phenotype-genotype (Table 1). Studies had to provide raw data that could be used to populate a 3x3 or 2x2 contingency table (see Tables 2 and 3) to allow one or measures of accuracy such as sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), or concordance to be calculated. Studies were not restricted based on age, disease group, or language.

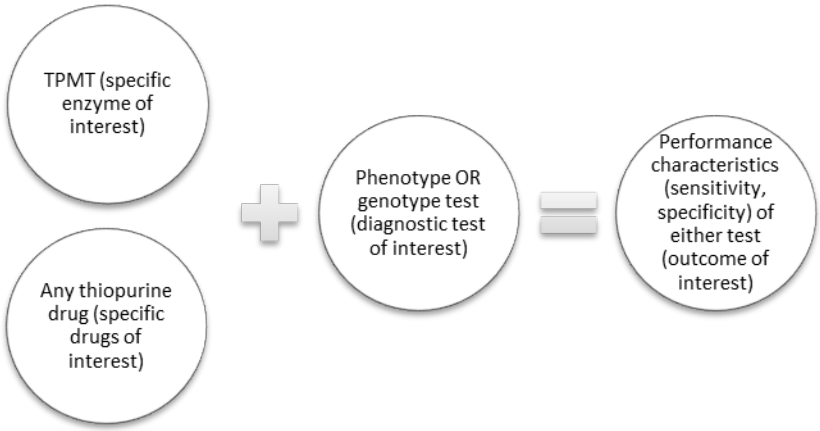
Table 1. Inclusion and exclusion criteria

Inclusion Criteria
Studies that evaluated either a TPMT genotype or TPMT phenotype technology in comparison to a reference standard.
Studies that 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.
Studies conducted in any age group
Studies conducted in any disease group
Studies published in any language, so long as it was possible to obtain sufficient translation to determine eligibility.
Exclusion Criteria
Studies not conducted in humans, including animal, tissue and *in vitro* studies.

2.1.2 Literature search design

A detailed search design ensured a comprehensive review. A conceptual framework was created (Figure 1) outlining the major concepts of the review question and the eligibility criteria. These concepts included TPMT, thiopurine drugs, phenotype testing or genotype testing for any thiopurine drug, and accuracy (sensitivity, specificity, positive and NPV, concordance). This framework provided the basis for developing preliminary search terms based on MeSH headings and known synonyms for each of these three components of the search. Medline and Embase, two major healthcare databases, were used in the design and testing of the search strategy.

Figure 2. Search strategy concept map



2.1.2.1 Search sources

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.

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) (Appendix B Grey literature sources).

2.1.2.2 Search term selection and development of search strategies

Search strategies were developed and terms were selected for each database in collaboration with a librarian experienced in systematic reviews and an experienced health technology assessment research team. Individual searches of databases were tailored to each database's subject/keywords terms. A protocol with a master list of MeSH terms and key words for textword searching derived from Medline was created as a guideline for developing a search strategy specific to each database. Search terms from the master list were entered into each database to identify relevant subject headings. Terms without subject headings were entered as textword search terms. The search protocol ensured systematic use of search terms and database-specific search entries. The search strategies for each citation database are found in Appendix C.

For initial development of the search strategies, common terms relating to each search concept, such as thiopurine drug names, were entered into the search field of the database. Medline search trees were explored for relevant MeSH topics and terms and Embase trees were searched for possible analogous terms. Any additional words deemed relevant based on clinical or research judgment were included. Groups of search terms were tested in Medline for relevance and the top 50 results were scanned for relevancy. If the addition of a search term did not improve the search results, the term was not included.

The search design was expanded to include thiopurine drugs in case relevant studies were not assigned MeSH terms relating to the TPMT enzyme. 'Methyltransferases' is a broader term than 'thiopurine methyltransferase', however it was decided to include the former term in the event that relevant studies had been assigned the broader MeSH term. The term was not exploded as search permutations revealed that exploding the term did not improve the specificity of the search results. Also, there are other enzymes down the tree from 'methyltransferases' that are not metabolizers of thiopurines and were therefore not included.

Regarding drug terms, all generic thiopurine drugs were listed in the search. MeSH terms for thiopurine drugs (such as their chemical derivatives) were not exploded, as this created excess noise (>9000 hits), and all relevant sub-groups were separately identified and included. Table 4

contains the master list of search terms and related terms from Medline used to develop search strategies for individual databases.

Table 2. Search terms and variations

Search term	Related terms
Thiopurine s-methyltransferase	Methyltransferases, thiopurine methyltransferase, thiopurine s methyltransferase, e.c. 2.1.1.67 †
Azathioprine	Azathioprine, thiopurine, azathioprine, imurel, immuran, imuran, 6 (1 methyl 4 nitro 5 imidazolyl) mercaptopurine, arathioprine, , aza-q, azafalk, azahexal, azamedac, azamun*, azanin, azapin, azapress, azaprine, azarex, azasan, azathiodura, azathipine, azathiopurine, azathropsin, azatioprina, azatrox, azatrilem, azopi, azoran, azothioprin, bw 57 322, bw57322, bw57-322, bw 57322, colinsan , immuther, imunen, imuprin, imurane, imurek, imuren, nsc 39084, nsc39084, thioazeprine, thioprine, transimune, zytrim, imidazole, nitroimidazole
Thioguanine	Thioguanine, thiopurine, thioguanin, tioguan, 2 amino 6 purinethiol, tabloid, 6 thioguanin*, lanvis, 2 amino 6 mercaptopurine, 2 amino purine 6 thiol, 2 aminopurine 6 thiol, 6 mercaptoguanine, 6 thioquanine, nsc 752, nsc752, thioguanidine
6-mercaptopurine	6-Mercaptopurine, thiopurine*, 6-thiopurine, leupurin*, bw 57323h, 6 mercaptopurine, 6 thiohypoxanthine, 6-mercaptopurine, monohydrate puri-nethol, 6-thiohypoxanthine, bw 57 323h, puri nethol, purimethol, purinethol, bw 57-323h, mercaptopurina, 17-dihydro-6h-purine-6-thione mercaptopurine, 6 mercapto purine, 6 mercaptopurin, 6 mercaptopurin monohydrate, 6 mp, 6 purinethiol, 6 purinethiol hydrate, lassen, empurine, ismipur, leukerin, loulla, mercaleukin, mercaptopurin, mercapurene, mern, mycaptine, nsc 755, nsc755, purine 6 thiol, purine thiol (6) ,purinethiol, thiohypoxanthine, xaluprine
Phenotype test	Phenotype, chromatography, high pressure liquid/ or Clinical Laboratory Techniques/ or Clinical Chemistry Tests/ or Cytological Techniques/ or Hematologic Tests/ or Radioligand Assay/ or Chemistry Techniques, Analytical/ or Enzyme Assays/ or Clinical enzyme tests/ or Mass spectrometry/ or Tandem Mass Spectrometry/ or Chromatography/ or Chromatography, High Pressure Liquid/ or Chromatography, Liquid/ or ("high adj3 liquid adj3 chromatograph*" or

	hplc or "cytologic* techniq*" or "cytologic* technic*" or (haematolog* adj2 (test or tests or testing)) or (haematolog* adj2 (test or tests or testing)) or "blood adj1 (test or tests or testing)" or radioassay* or radioreceptor* or radioligand*).mp
Genotype test	Pharmacogenetics/ or Genetics/ or Genetic Testing/ or Heterozygote detection/ or Genotype/ or exp Sequence Analysis/ or Heterozygote/ or Homozygote/ or Hemizygote/ or genetic techniques/ or genetic association studies/ or genome-wide association study/ or genetic testing/ or genotyping techniques/ or Restriction Mapping/ or Polymorphism, Restriction Fragment Length/ or Polymorphism, Single-Stranded Conformational/ or Oligonucleotide Array Sequence Analysis/ or exp Polymerase Chain Reaction/ or Biomarkers, Pharmacological/ or Genetic Markers/ or Biological Markers/ or genetic predisposition to disease/ or ("Amplification Refractory Mutation System" or microchip or pharmacogen* or toxicogen* or ((genotyp* or genetic*) adj2 (test or tested or tests or testing or predispos* or screen*)) or biomarker* or PCR).mp.
Diagnostic accuracy	Evaluation Studies, comparative study, validation studies, sensitivity, specificity, predictive value of tests, diagnostic errors, false negative reactions, false positive reactions, observer variation, positive predictive value, negative predictive value, accuracy

† chemical identifier for thiopurine s-methyltransferase

2.1.2.3 Search permutations

Search terms were as detailed as possible, and there were several possible ways to combine search concepts, making preliminary search results quite broad. Assessment of the first 50 results of the initial search in Medline revealed that it was difficult to determine relevance without a detailed review. As such, a specificity check to eliminate excessive false positives was devised using the results of a previous TASK systematic review of TPMT testing as a guide [26]. For the specificity check, 24 different search permutations were constructed and run in Medline to test which combination was most specific and sensitive, and results were exported to EndNote. Permutations combined, for example, terms such as phenotype testing and genotype testing using ‘and’, then ‘or’, or first searching phenotype testing or genotype testing before combining with accuracy terms (e.g. sensitivity, specificity). One reviewer (LR) then determined how many of the previously identified studies were included in the results of each search permutation.

Through this process, the research team was able to have a valid reference to determine whether the search strategy was specific and sensitive to the research question.

It was hypothesized that the permutation combining phenotype and genotype in the search ('phenotype AND genotype') would be the most specific, but this was found to be the least specific permutation (9/17 articles missing). Several other permutations were somewhat better, but had significantly higher numbers of total search results (for instance, "phenotype OR genotype" had n ~ 6000 hits). Consideration was given to as the fact that screening greater numbers of papers would introduce greater judgment into the selection process and consequently increase the risk of bias.

It was also determined that the use of the search terms related to accuracy (e.g. sensitivity, specificity) did not improve search permutations as these terms were not commonly assigned MeSH terms in Medline and consequently, produced significant amounts of noise and reduced specificity.

Consideration was also given to the potential to miss articles from narrower searches, the possibility of reducing sensitivity with searches in which concepts were removed (such as removing terms related to accuracy), and the high volume of papers for screening if the search was conducted with high sensitivity. The research team determined that the final search permutation with the least number of excess hits and with the most number of known relevant articles would be used.

2.1.2.4 Final search strategy

The most comprehensive search strategy combined the search concepts in the following manner: TPMT (or related terms) AND a thiopurine drug (common thiopurine drugs such as AZA, 6-mercaptopurine (6-MP), and thioguanine) AND either a phenotype OR genotype technology. This combination of terms maintained relatively high specificity for well-known studies, with 16/17 of previously identified studies detected in the results. Detailed search strategies for Medline and other citation databases are available in Appendix A. Grey literature sources are listed in Appendix B.

2.1.3 Review for eligibility

Two reviewers (RZ and LR) 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 (WU) as needed. Agreement became consistent after comparing categorization of approximately 130 abstracts and titles between the two reviewers. Subsequently, one reviewer (LR) screened the remaining titles and abstracts.

A reference manager software program (EndNote Program X4) was used to maintain reference citations. Search results from each citation database were imported into a single EndNote Library. At the time they were imported, all results were labeled with the database from which they were retrieved. Relevant grey literature was entered manually. Duplicates were subsequently removed based on matching journal volume and page range rather than title or author, as the latter fields can vary slightly between citation databases. After the removal of duplicates, electronic folders were created within EndNote for each category of screening (‘yes’, ‘no’, and ‘maybe’) and all references were categorized. Folders for ‘yes’, ‘no’, and ‘maybe’ were created for each reviewer so that reviewers could independently assign categories.

The full texts of abstracts and titles categorized as ’maybe’ and ‘yes’ were retrieved through the University of Toronto Library Services catalogue, requested through an inter-library loan or requested directly from the author. Only 3/67 interlibrary loan requests were not fulfilled. ’Maybe’ citations were reviewed in full text by one reviewer (LR) for inclusion and studies with inconclusive eligibility were reviewed in full text by a second reviewer (RZ). A second reviewer (RZ) reviewed the full text of 5% of all included studies as a quality control check. If at any point there was incongruence between reviewers, discrepancies were resolved through discussion with decision rules generated where necessary.

Inclusion and exclusion criteria were revisited after reviewers began reviewing search results, as some studies reported ADE rates in comparison to the frequency of positive test results. Investigators considered the option of including ADE rates as a measure of test accuracy. Since ADEs do not depend exclusively on the result of the diagnostic test, but also depend on the

thiopurine dose selected following the diagnostic test result, it was decided that ADE rate would not be a useful independent measure of diagnostic test accuracy.

2.1.3.1 Translation

Translation was required for papers published in Chinese, Dutch, French, German, Japanese, Korean, Polish, Serbian, and Spanish. Local volunteers were recruited and offered a small honorarium. Volunteers screened articles based on study purpose and inclusion/exclusion criteria. Verbal explanation was provided where requested and in cases where clarification was needed, in person sessions with a member of the research team was arranged.

Non-English publications that met inclusion criteria were reviewed in full with the translator to facilitate understanding of data collection process. Where multiple articles required translation (e.g. Mandarin), the translator completed the data extraction for two studies alongside one reviewer, and completed the rest independently. All extracted data from these remaining papers were reviewed by a member of the research team to clarify missing or unclear items.

2.1.4 Data extraction

A data extraction tool was created using Microsoft Access (version 2010) to ensure consistent abstraction of relevant data from each study. The elements of the tool used for data extraction are included in Appendix C. Data abstracted included study design characteristics, such as epidemiological design, primary objective, inclusion and exclusion criteria, recruitment time frame, recruitment strategy, year of study, target gender, and target race or ethnic group. Data were also collected describing the study sample, including number of subjects recruited, number of subjects tested, mean/median age or age range as reported, the proportion of male subjects, sample race or ethnic groups reported, number of subjects with target disease, and relationship of subjects (related or not related to one another). Data were collected describing test characteristics, such as blood component collected, phenotype method, substrate used, cutpoints used, and whether cutpoints were previously validated.

If test performance results including sensitivity, specificity, PPV, NPV, and concordance were reported, then they were abstracted as reported by the authors. In addition, 2x2 or 3x3 contingency tables were populated for each included study using raw data reported in tables, text or inferred from graphs, making it possible for the reviewers to calculate test performance characterstics independently. Templates of contingency tables are found in Tables 3, 4, and 5.

As no gold standard reference test exists, the phenotype TPMT test was established as the reference test and the genotype as the index test, for all calculations of sensitivity and specificity, since the latter represents the innovative technology. By designating the phenotype TPMT test as the reference standard for the calculated sensitivity and specificity regardless of the reference test reported by authors, a standardized approach to analysis could be undertaken. This facilitates the comparison between studies and prepares the data for meta-analysis (MA) in the next phase of the research.

Studies used terminology inconsistently, sometimes using ‘low’ and ‘intermediate’ to represent the same level of activity; likewise ‘deficient’ and ‘absent’ were used interchangeably. For the purpose of this report ‘absent’ and ‘deficient’ activity were considered equivalent to, and were classified in, the category of ‘low’ enzyme activity. The term ‘intermediate’ was used to describe intermediate enzyme activity. The terms ‘high’ activity and ‘normal’ were both interpreted to represent the upper spectrum of enzyme activity, which was categorized as ‘high/normal’ (presumed wild type genotype).

For each included paper providing sufficient data, reviewers first classified TMPT activity into 3x3 tables after considering the cutpoints reported by the study author, the distribution of the TPMT activity results (e.g. graphical distributions provided in the text by the study author), and the description provided in the text (Table 3). ‘Low’ activity included reported ‘deficient’ or ‘absent’ activity, or enzymatic activity below approximately 5 U/ml packed red blood cells (pRBC). Reported activity above 5 U/ml pRBC and below approximately 10 U/ml pRBC was categorized as ‘intermediate’ activity. Enzymatic activity reported above 10 U/ml pRBC was classified as ‘high/normal’. These activity levels reflect a common classification of TPMT activity which was initially reported by Weinshilboum et al.[8].

Table 3. Contingency table (3x3) template for detection of homozygous mutation

	Low Enzyme Activity	Intermediate Enzyme Activity	High/Normal Enzyme Activity	
Homozygous Mutation				Total persons with homozygous mutation
Heterozygous Mutation				Total persons with heterozygous mutation
Wild-Type				Total persons with wild-type (mutations absent)
	Total persons with low enzyme activity	Total persons with intermediate enzyme activity	Total persons with high/normal enzyme activity	Total persons tested

Phenotype is designated as the reference test.

Study authors did not always provide sufficient data to populate all the cells in a 3x3 contingency table. Where 3x3 tables were not possible, 2x2 cells were populated. A 2x2 table was also created from the studies with 3x3 tables by combining the categories to create a 2x2 table based on the data reported. In cases where data were not explicitly described, decision rules were applied by reviewers to guide the cell assignment. There were two alternatives for collapsing a 3x3 table; the first was to define ‘low’ and ‘deficient’ enzyme activity as the presence of a homozygous mutation. An example of this 2x2 is shown in Table 4. This results in grouping intermediate and high enzyme activity together, and grouping heterozygous and wild-type genetic expression together.

Table 4. Contingency table (2x2) template for detection of homozygous mutation

	Test positive for low enzyme activity	**Test negative for low enzyme activity (intermediate + high activity)**	
Homozygous mutation present			Total persons with homozygous mutation present
Homozygous mutation absent (heterozygous + wild-type)			Total persons with homozygous mutations absent
	Total persons with positive test for low enzyme activity	Total persons with negative test for low enzyme activity	Total persons tested

Phenotype is designated as the reference test

The second option was to define test performance based on the presence of any mutation (heterozygous or homozygous) versus the absence of any mutation (wild-type). Reviewers calculated sensitivity and specificity for both of these approaches due to their different clinical diagnostic implications. Table 5 shows a 2x2 table with the categories reflecting this approach.

Table 5. Contingency table (2x2) template for detection of homozygous or heterozygous mutation

	Test positive for low + intermediate enzyme activity	**Test negative for low + intermediate enzyme activity (High enzyme activity)**	
Mutation present (homozygous + heterozygous)			Total persons with mutations present
Mutation absent (wild-type)			Total persons with mutations absent
	Total persons with low + intermediate enzyme activity	Total persons with high enzyme activity	Total persons tested

Phenotype is designated as the reference test

2.2 Quality appraisal

The Quality Assessment tool for Diagnostic Accuracy Studies (QUADAS) version 2 was used to evaluate quality of the included studies. The QUADAS-2 contains four domains pertaining to i) patient selection, ii) the index test, iii) the reference standard, and iv) flow and timing of the study. The QUADAS-2 ascertains the risk of bias associated with each of the four domains, as well as the applicability of the first three domains to the research question by asking specific questions about bias and applicability. The tool allows items within standard domains to be added or modified by reviewers. A fifth domain, as described below, was created for the purpose of this study to assess the risk of bias pertaining specifically to genomic tests.

A study-specific QUADAS-2 appraisal tool was created using Microsoft Access (version 2010) by tailoring items to the specific study objectives to ensure consistent and reliable assessment between reviewers (Appendix D). The tool was piloted with two reviewers (RZ, LR) for usability, readability and relevance with two well-known papers [9, 15] established as relevant in a previous literature review [26]. Discussion between reviewers resolved discrepancies and any outstanding disagreements were discussed with the principal investigator (WU). As a result of the pilot appraisal, several revisions of the QUADAS-2 data collection tool were made to improve clarity and organization.

The QUADAS-2 provides a flexible approach to evaluating quality by providing a framework open to adding relevant questions to the domains. Given the context of genomic diagnostic tests, several issues relevant to genomic testing were not directly captured by the existing domains. For example, selectively genotyping individuals based on their phenotype test result (e.g. genotyping only those with intermediate or low phenotype) would bias sensitivity and specificity calculations, and was therefore thought to be an essential component to consider. As such, the investigators added a domain using the open-ended model provided by the QUADAS-2 [27]. A domain was created reflecting characteristics of genetic diagnostic testing that were absent from the QUADAS-2 tool. The genomics domain was piloted with two reviewers (RZ, LR) until consensus was reached regarding the design and application to quality appraisal. The domain is presented in Table 6 (see also Appendix C).

Table 6. Fifth domain created to address risk of bias with genomic technologies

Domain 5 – Genomics	
Question	**Response options**
❖ Did the study clearly identify the racial or ethnic population being studied?	Yes/No/Unclear
❖ If specific polymorphisms were tested, were they identified a priori?	Yes/No/Unclear
❖ Was the classification of genetic test results (homozygous/heterozygous/wild type or genetic score) described clearly?	Yes/No/Unclear
❖ Did the study avoid selectively genotyping after establishing phenotype (or vice versa)?	Yes/No/Unclear
❖ Were all polymorphisms of interest tested?	Yes/No/Unclear
*RISK: Could genomic-specific process have introduced bias?	RISK: Low/High/Unclear

The QUADAS-2 was completed by answering selected questions pertaining to each of the five domains. Each domain had a descriptive component, followed by items pertaining to bias and applicability. For both bias and applicability, the reviewer determined whether the study had 'high bias or concern for applicability', 'low bias or concern for applicability', or whether the study was 'unclear'. Decision rules were generated for interpretation of each QUADAS-2 question and item in each domain and revisited frequently to ensure consistent judgment of bias and applicability throughout the quality appraisal. Criteria for determining when a domain as a whole should be considered to have 'Low bias or concern', 'High bias or concern' or 'Unclear' were established (see appendix E).

Two reviewers (LR, RZ) independently appraised the first few papers.Consensus was reached after four papers and one reviewer (LR) completed the quality appraisal for the remaining papers. The second reviewer completed a QUADAS-2 quality appraisal independently for 5% of all remaining studies as a quality control measure, as well as for any papers that were uncertain to the first reviewer.

2.2.1 Decision criteria for determination of study quality

An overall determination of high versus low quality of included studies was made based on a clear, pre-established algorithm (Table 7) created by the reviewers and reviewed for consistency until consensus was reached. When necessary, previous studies were revisited to ensure consistency.

Studies were considered to be of 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 considered low quality. If only one domain demonstrated high risk of bias, then the study was considered to be of high quality overall. If the study had two or more domains that were uncertain, then the study was deemed as low quality overall.

Table 7. Decision criteria for high/low quality studies

QUADAS Results – L (low), H (high), or U (unclear) for each domain		Quality	Rationale
Risk of Bias (5 domains)	**Concern for Applicability (3 domains)**		
LLLLL	LLL	High	All domains have low bias and low concern for applicability
HHHHH	HHH	Low	All domains have high bias and high concern for applicability
H ≥ 2	H ≥ 2	Low	The number of domains with high bias OR with high concern for applicability is two or more
U ≥ 2	U ≥ 2	Low	There are two or more domains with unclear risk or unclear applicability concerns

Abbreviations: L = Low; H = High; U = Unclear

3 RESULTS

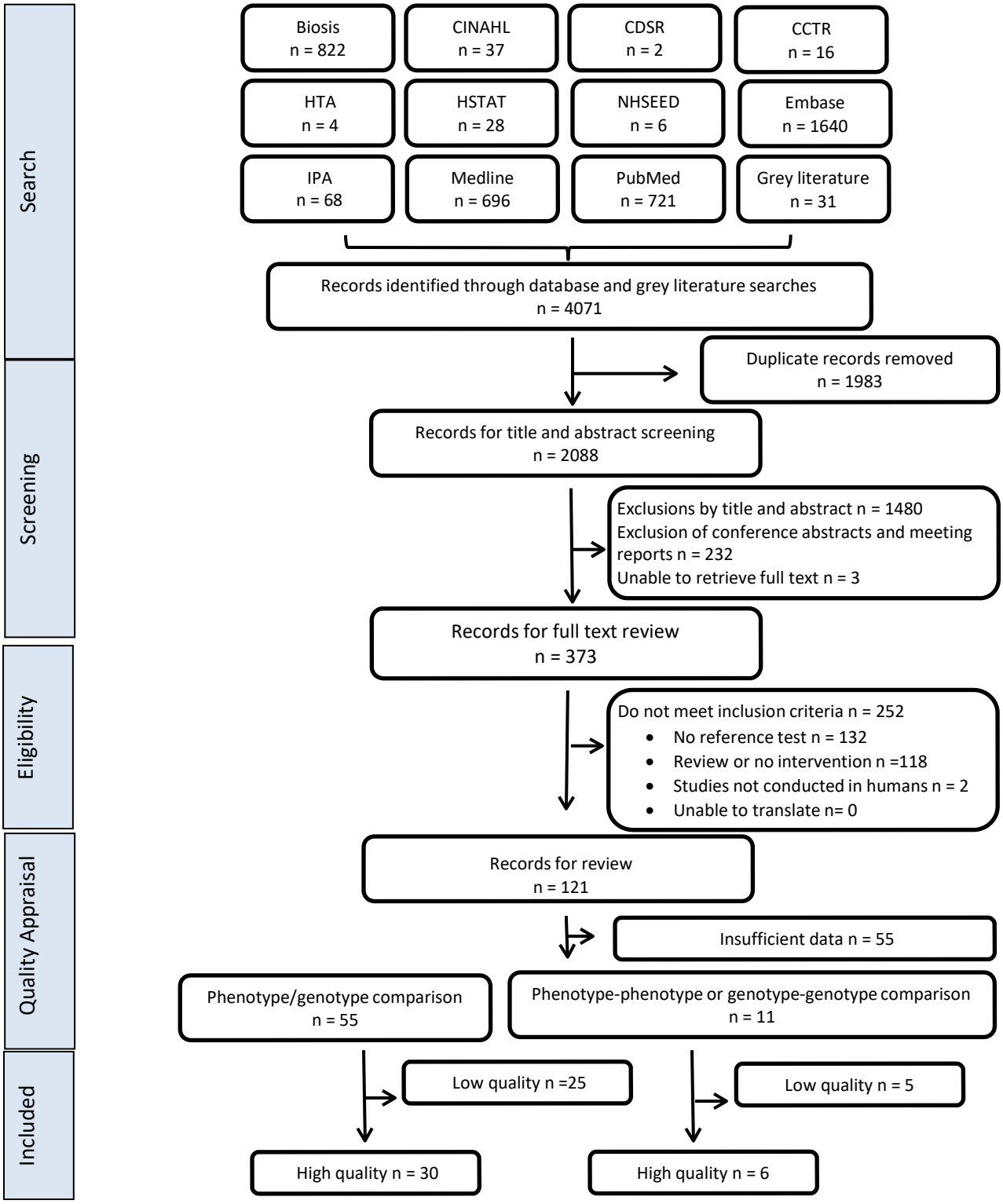
The following sections describe the final search results, the characteristics of the studies identified through the search, the results of the quality appraisal for phenotype-genotype, phenotype-phenotype, and genotype-genotype comparisons, detailed study characterstics of the high quality studies and the different TPMT testing methods.

3.1 Systematic review

3.1.1 Quantity of publications

The search results are displayed in as PRISMA flow chart in Figure 2. The final search yielded 4071 publications from the citation database and grey literature sources. After the removal of duplicates, 2088 records were screened for inclusion. First, titles and abstracts were screened, resulting in 374 full text papers to be screened, of which 37 required translations from Korean, German, Polish, French, Japanese, Chinese, Dutch, Spanish, and Serbian. One hundred and twenty-one papers appeared to meet inclusion criteria and were assessed for relevant data. Of these, 55 had insufficient data to answer the review question, and were excluded from the review.

Figure 3. PRISMA flowchart



3.1.2 Study characteristics of eligible papers

Sixty-six papers contained sufficient data to address the review question, of which 55 reported a phenotype-genotype comparison [9-16, 28-74]. The remaining 11 papers reported a laboratory method comparison (either phenotype–phenotype, or genotype-genotype) [11, 12, 74-82]. All 66 were carried forward for quality appraisal.

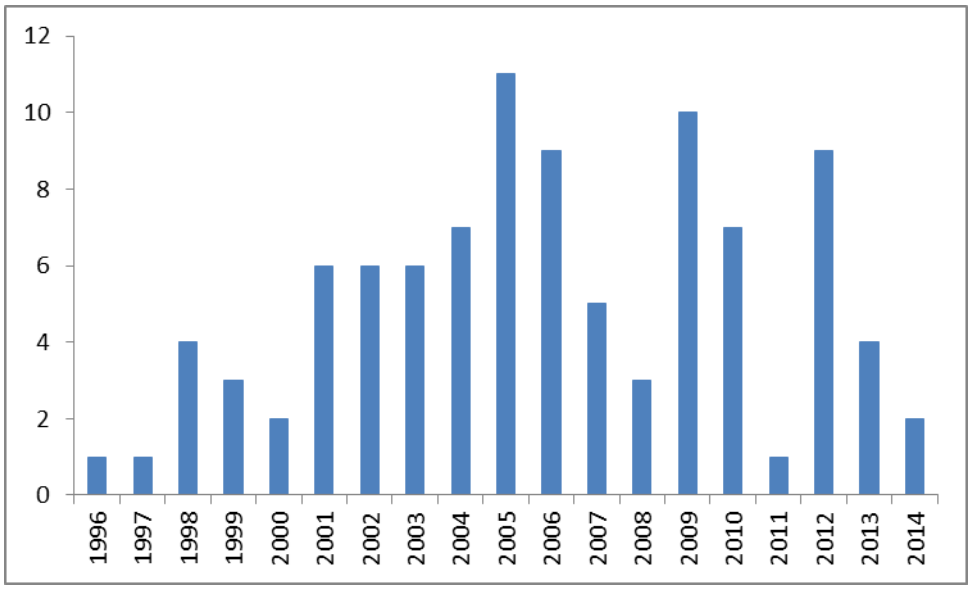
Studies comparing phenotype and genotype testing were published between 1996 and 2014. Annual trends are presented in Figure 3. Studies comparing phenotype-phenotype or genotype-genotype were published between 1994 and 2013.

Among the 66 eligible studies, sample sizes ranged from 15 [65] to 7195 [36]. Sixteen studies were conducted in adults, 11 in children, 13 in a mix of adult and pediatric populations, and the remaining 26 did not specify the age of the study sample. Fourteen studies were conducted in healthy populations while 51 studies sampled patients. These included 14 with ALL patients, 15 in IBD, six were not specified, 13 were 'other' patients, one was dermatological conditions, and two were in organ transplant patients. Of these 66 studies, 17 contained a mix of different disease populations. One study did not specify the disease population [75]. Many studies identified a particular ethnicity, race or nationality, including Caucasian (n=11), Chinese (n=4), European (n=5) and German (n=1). The 'other' category (n=19) included Estonians, Portuguese, Spanish, Chilean, Italian, New Zealand, Scandinavian, Tunisian, Alaskan, South African, Bulgarian, Jewish, Japanese, Malays, British and Irish. Authors did not commonly identify whether participants were related to one another; only 18 studies reported that participants were unrelated. The remaining did not specify this information.

With regard to laboratory test methods, the most commonly reported genotype amplification measure was polymerase chain reaction (PCR) (n=29), followed by allele-specific PCR (AS-PCR) (n=11). PCR was reported in conjunction with single strand conformation polymorphism (PCR-SSCP) in two studies [29, 74]. Other methods of genotyping reported included restriction fragment length polymorphism (RFLP) (n=19), denaturing high performance liquid chromatography (DHPLC) (n=4), direct sequencing, pyrosequencing, SNaPshot sequencing (n=8), multiplex amplification refractory mutation (ARMS) (n=3), TaqMan single nucleotide polymorphism (SNP) Genotyping (n=3), LightSNiP genotyping (n=1) and matrix-assisted laser

desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) (n=1). DHPLC, although more commonly used in phenotype testing, was used in four studies. Five studies did not report the amplification method used [36, 45, 72, 73, 82], and one study did not specify a genotype method at all [73]. Several studies reported more than one method of genotyping, which sometimes varied depending on the SNP being tested. Phenotype methods used were generally either high performance liquid chromatography (HPLC) (n=22) or radiochemical (RC) method (n=19). Although rare, mass spectrometry (n=2) and competitive micro-well immunoassay (n=1) were also reported, while 11 studies failed to clearly specify the phenotype method used.

Figure 4. Number of phenotype-genotype publications per year



3.2 Quality appraisal

3.2.1 Phenotype-genotype comparisons

The 55 papers with sufficient data to calculate sensitivity and specificity for the test of interest were appraised for quality. Of these, 30 studies were deemed of high quality. The remaining 25 had insufficient or low quality information reported pertaining to the recruitment of participants, conduct of the tests, flow and timing of the study, and reporting of genomics-related aspects. Seven of the 55 studies demonstrated ‘high’ or ’unclear’ concern regarding applicability to the review question for at least one of the five domains.

Fifteen of 30 high quality studies showed ‘high’ or ’unclear’ risk of bias for at least one of the five domains. Thirteen of the studies consistently demonstrated low scores (low risk of bias, low concern for applicability) (Table 8). Low quality studies generally had more ‘unclear’ ratings than high quality studies, as opposed to definitive ‘high’ risk ratings. Only nine low quality studies were deemed of low quality due to two or more ‘high’ risk or concern for applicability in the absence of ‘unclear’ ratings. The remaining 16 studies had at least one element that was considered ‘unclear’ in addition to one or more elements of ‘high’ or ‘unclear’ risk or concern for applicability (Table 9).

Among the 25 low quality studies, the highest risk of bias was observed for Domain 4 (Genomics), with 12 studies appraised as ‘high’ bias. Risk of bias was next most frequent in Domain 3 (Reference test), with seven studies appraised as ‘high’ bias. High bias was observed for six studies for Domain 4 (Flow & timing), 5 studies for Domain 2 (Index test), and four studies for Domain 1 (Patient selection). For the domains reporting ‘unclear’ risk of bias or applicability, the most problematic domain was Domain 3 (Reference test) with 12 studies having insufficient information to determine whether bias was high or low. Seven studies were ‘unclear’ for Domain1 (Patient selection), seven studies for Domain 5 (Genomics), five studies for Domain 4 (Flow & Timing) and one study for Domain 2 (Index test). Concern for applicability was highest in Domain 3.

Table 8. QUADAS-2 results for high quality phenotype-genotype studies

Author	Year	Domain 1 - Patient Selection		Domain 2 - Index Test		Domain 3 - Reference test		Domain 4 - Flow and Timing	Domain 5 - Genomics
		Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Risk of bias
Ben Salah [30]	2013	Low	Low	Low	Low	Unclear	Low	Low	Low
Fakhoury [31]	2007	Low	Low	Low	Low	Low	Low	Low	Low
Fangbin [32]	2012	Low	Low	Low	Low	High	Low	Low	Low
Ford [33]	2006	Low	Low	Low	Low	High	Low	Low	Low
Ford [34]	2009	Low	Low	Low	Low	Low	Low	Unclear	Low
Ganiere-Monteil [66]	2004	Low	Low	Low	Low	High	Low	Low	Low
Gazouli [35]	2012	Low	Low	Low	Low	Low	Low	Low	Low
Hindorf [36]	2012	Low	Low	Low	Low	Unclear	Low	Low	Low
Jorquera [46]	2012	Low	Low	Low	Low	Low	Low	Low	Low
Langley [67]	2002	Low	Low	High	Low	Low	Low	Low	Low
Larussa [37]	2012	Low	Low	Low	Low	Low	Low	Low	Low
Lennard [38]	2012	Low	Low	Low	Low	Low	Low	Low	Low
Liang [39]	2013	Low	Low	Low	Low	Low	Low	Low	Low
Loennechen [28]	2001	Low	Low	Low	Low	Low	Low	Low	Low
Ma [47]	2006	Low	Low	Low	Low	Unclear	Low	Low	Low
Marinaki [68]	2003	Low	Low	Low	Low	Unclear	Unclear	Low	Low

Author	Year	Domain 1 - Patient Selection		Domain 2 - Index Test		Domain 3 - Reference test		Domain 4 - Flow and Timing	Domain 5 - Genomics
		Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Risk of bias
Milek [40]	2006	Low	Low	Low	Low	Low	Low	Low	Low
Oselin [41]	2006	Low	Low	Low	Low	High	Low	Low	Low
Schaeffeler [70]	2004	Low	Low	Low	Low	Low	Low	Low	Low
Schwab [72]	2002	Low	Low	Low	Low	Low	Low	Low	Low
Serpe [42]	2009	Low	Low	Low	Low	Unclear	Low	Low	Low
Spire-Vayron de la Moureyre [74]	1998	Unclear	Low	Low	Low	Low	Low	Low	Low
Spire-Vayron de la Moureyre [29]	1998	Unclear	Low	Low	Low	Low	Low	Low	Low
von Ahsen [73]	2005	Low	Low	Low	Unclear	Unclear	Low	Low	Low
Wennerstrand [45]	2013	Low	Low	Low	Low	Low	Low	Low	Low
Winter [15]	2007	Low	Low	Low	Low	Low	Low	Low	Low
Wusk [16]	2004	Low	Low	Low	Low	High	Low	Low	Low
Xin [43]	2009	Low	Low	Low	Low	Low	Low	Low	Low
Yates [9]	1997	Low	Low	Low	Low	Low	Low	Low	High
Zhang [44]	2007	Low	Low	Low	Low	Unclear	Low	Low	Low

Table 9. QUADAS-2 results for low quality phenotype-genotype studies

Author	Year	Domain 1 - Patient Selection		Domain 2 - Index Test		Domain 3 - Reference test		Domain 4 - Flow and Timing	Domain 5 Genomics
		Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Risk of bias
Alves [10]	2001	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Anglicheau [11]	2002	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Ansari [48]	2002	High	Low	Low	Low	Low	Low	Low	High
Arenas [49]	2005	Unclear	Low	Low	Low	Low	Low	Low	Unclear
Barlow [50]	2010	Low	Low	Low	Low	Unclear	Low	High	High
Ebbesen [51]	2013	High	Low	High	Low	High	Low	Low	High
el-Azhary [65]	2009	Unclear	Low	High	Low	Unclear	Unclear	Unclear	Unclear
Evans [52]	2001	High	Low	Low	Low	Low	Low	Low	High
Ferucci [53]	2011	Low	Low	Unclear	Unclear	Unclear	Low	Low	High
Gu [54]	2003	Unclear	Low	Low	Low	Unclear	Low	Unclear	Low
Haglund [55]	2004	High	Low	Low	Low	Low	Low	Low	Unclear
Heckman [56]	2005	Low	Low	Low	Low	High	Low	Unclear	Unclear
Hon [57]	1999	Low	Low	Low	Low	Low	Low	High	High
Indjova [12]	2003	Low	Low	Low	Low	Low	Low	High	High
Kasirer [58]	2014	Low	Low	Low	Low	High	Low	Low	High
Kow Yin Kham [13]	2008	Low	Low	High	High	High	Low	Low	High
Larovere [14]	2003	Low	Low	Low	Low	Unclear	Low	Low	High
Reis [59]	2003	Low	Low	Low	Low	High	Low	High	High

Author	Year	Domain 1 - Patient Selection		Domain 2 - Index Test		Domain 3 - Reference test		Domain 4 - Flow and Timing	Domain 5 Genomics
		Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Risk of bias
Relling [69]	1999	Unclear	Low	Low	Low	High	Low	Unclear	Unclear
Relling [60]	1999	Low	Low	Low	Low	High	High	High	Low
Rossi [61]	2001	Low	Low	Low	Low	Unclear	Low	Low	Low
Schmiegelow [62]	2009	Low	Low	High	Low	Unclear	Low	Low	Unclear
Schmiegelow [71]	2009	Low	Low	High	Low	Unclear	Unclear	High	Low
Sies [63]	2005	Low	Low	Low	Low	Unclear	Low	Low	High
Tamm [64]	2008	Unclear	Low	Low	Low	Unclear	Low	Unclear	Unclear

3.2.2 Genotype-genotype and phenotype-phenotype comparisons

For the studies which compared genotype-genotype or phenotype-phenotype testing, 11 studies had sufficient data for extraction of which six were found to be of high quality. The QUADAS-2 results for high quality and low quality studies are presented in Table 10 and Table 11, respectively.

Of the six high quality studies, all were genotype-genotype test comparisons. Three of these studies had ‘unclear’ risk of bias in Domain 1 (Patient Selection) [78, 81, 83]. Lu et al. and Ma et al. had ‘unclear’ risk of bias in Domain 3 (Reference Test) and Domain 5 (Genomics), respectively [76, 77]. One study had ‘low’ risk of bias and ‘low’ concern for applicability in all domains [82].

The five low quality studies did not have clear patterns of bias risk or concern for applicability. Two of three low quality genotype-genotype studies indicated high bias in Domain 5 (Genomics). Three studies had ‘unclear’ risk of bias in Domain 1 (Patient selection), two studies had ‘unclear’ risk of bias in Domain 2 (Index test), two studies had ‘unclear’ risk of bias in Domain 3, two studies had ‘unclear’ concern for applicability in Domain 3 (Reference test), and one study had ‘unclear’ risk of bias in Domain 4 (Flow and timing). For Domain 3 (Reference test), two studies were ‘unclear’ as to whether there was concern regarding applicability to the research question.

Two studies had low risk of bias in Domain 1, and all five had low concern for applicability. Three studies had low risk of bias in Domain 2, and all studies had low concern for applicability. Three studies had low risk of bias in Domain 3, while only three of five had low concern for applicability. Four of five studies had low risk of bias for Domain 4, and only one study had low risk of bias in Domain 5.

Two studies were phenotype-phenotype studies, and therefore Domain 5 (Genomics) did not apply, and could not have a risk of bias reported.

Table 10. QUADAS-2 results for high quality phenotype-phenotype or genotype-genotype studies

Author	Year	Domain 1 - Patient Selection		Domain 2 - Index Test		Domain 3 - Reference test		Domain 4 - Flow and Timing	Domain 5 - Genomics
		Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Risk of bias
Chowdury [78]	2007	Unclear	Low	Low	Low	Low	Low	Low	Low
Kim [83]	2013	Unclear	Low	Low	Low	Low	Low	Low	Low
Lu [84]	2005	Low	Low	Low	Low	Unclear	Low	Low	Low
Ma [77]	2003	Low	Low	Low	Low	Low	Low	Low	Unclear
Roman [81]	2012	Unclear	Low	Low	Low	Low	Low	Low	Low
Schaeffeler [82]	2008	Low	Low	Low	Low	Low	Low	Low	Low

Table 11. QUADAS-2 results for low quality phenotype-phenotype or genotype-genotype studies

Author	Year	Domain 1 - Patient Selection		Domain 2 - Index Test		Domain 3 - Reference test		Domain 4 - Flow and Timing	Domain 5 - Genomics
		Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Concern for applicability	Risk of bias	Risk of bias
Anglicheau [11]	2002	Low	Low	Unclear	Low	Low	Unclear	Low	n/a
Indjova [85]	2003	Unclear	Low	Low	Low	Low	Low	Low	High
Lennard [75]	1994	Unclear	Low	Unclear	Low	Unclear	Low	Low	n/a
Osaki [80]	2011	Low	Low	Low	Low	Unclear	Unclear	Low	Low
Spire-Vayron de la Moureyre [74]	1998	Unclear	Low	Low	Low	Low	Low	Unclear	High

3.3 Design characteristics of high quality studies

3.3.1 Study objectives and eligibility criteria

High quality phenotype-genotype studies were published between 1997 and 2013. The primary objectives and eligibility criteria for the high quality phenotype-genotype comparisons are listed in Table 12. Eleven studies stated their primary objective was to investigate the relationship (e.g. concordance) between phenotype and genotype testing for TPMT activity determination [16, 31, 33-36, 38, 39, 44, 47, 70]. Two studies explicitly stated that investigating this relationship was a secondary objective [29, 41].

Inclusion criteria for the individual studies often specified that participants should meet specific disease criteria: healthy [40-42], acute lymphocytic leukemia (ALL) [31, 38, 45], IBD [16, 35, 72, 73], transplant [39, 43], and renal failure [44]. One study specified pediatric patients in their inclusion criteria [31]. Common exclusion criteria were a history of blood transfusions [32, 37, 38], concurrent medications such as methotrexate [32, 39, 66], insufficient functioning of major organs [32], and concurrent or history of a variety of acute, chronic or genetic diseases [39, 42, 44, 66]. One study specified that blood samples more than 8 days old were excluded [33]. Most studies did not specify exclusion criteria [9, 15, 16, 28-31, 34-36, 40, 41, 43, 45-47, 68, 70, 72-74].

Recruitment of patients and conduct of studies ranged from a 4-week period to over 7 years. Most studies did not specify the time period during which patients were recruited.

For high quality genotype-genotype studies, no studies reported that the primary objective was to measure sensitivity and specificity (Table 13). However, all studies used general terms implying an intent to evaluate one technology compared to another. Terms such as “compare”, “report on a new method”, “validate”, “investigate”, “establish and apply” were commonly used to describe the purpose of the study. One study did not report on the time frame of the study or the patient population [83]. Of the remaining studies, two reported a recruitment time frame of seven years and four years [77, 81], and all reported the patient population. Patient populations included children with ALL [77, 84] and patients with beta thalassemia [76], patients with IBD [78], and otherwise healthy volunteers [82, 84]. One study specified patients but did not specify which kind of patient they were [81].

Table 12. Design characteristics of high quality phenotype-genotype studies

Author	Year	Primary objective	Inclusion criteria	Exclusion criteria	Study period
Ben Salah	2013	Investigate TPMT activity distribution and allele frequency of common alleles	Not specified	Not specified	Not specified
Fakhoury	2007	Study correlations between TPMT genotype and enzyme activity	Children diagnosed with ALL; enrolled in two consecutive European trials	Not specified	Nov 1991 - Dec 2005 (14y)
Fangbin	2012	Role of phenotype and genotype in predicting leucopenia	Patients with steroid dependent disease, frequent relapses, on remission maintenance, & postoperative prophylaxis	Blood transfusions, cyclosporin, or methotrexate before treatment initiation; interfering treatments (e.g. allopurinol, diuretics); insufficient function of heart, liver, kidneys; active infection; pregnancy	Aug 2006 - Aug 2011
Ford	2006	Compare new method phenotype (whole blood) with old method (RBC lysate) and genotype	Routine samples collected over 4 wk period	Any samples >8 days old	4wk period (Nov 2005)
Ford	2009	Examine the phenotype-genotype concordance to investigate effectiveness as QA tool	All consecutive routinely collected samples	Not specified	July 2007 - July 2008
Ganiere-Monteil	2004	Investigate the impact of age on TPMT activity by comparing TPMT activity (pheno and geno) in healthy young Caucasians from birth (cord blood) to adolescence with adult	Patients with IBD; taking AZA or 6-MP for at least 3 months or experienced adverse events with these drugs; dose between 0.3-2.5mg/kg.	Past history of acute, chronic or genetic diseases; receiving medications	Not specified

Author	Year	Primary objective	Inclusion criteria	Exclusion criteria	Study period
		Caucasians			
Gazouli	2012	Examine sensitivity and specificity of TPMT genotyping for TPMT enzymatic activity	Patients with diagnosis of IBD; patients using AZA or 6-MP >3mo or adverse event during tx; dosage range specified	Not specified	Feb 2007 - Aug 2011
Hindorf	2012	Investigate the correlation between TPMT genotype and phenotype; analyze the results from a clinical and practical perspective	Unselected and consecutive TPMT phenotype and genotype determinations sent to the study site	Not specified	2006 - 2010
Jorquera	2012	Study the TPMT activity and genotype in Chilean subjects	Healthy persons; older than 18 years; unrelated	Not specified	Not specified
Langley	2002	Determine whether the phenotypes or genotypes correlate with clinical outcomes for AZA therapy	Patients attending the autoimmune liver disease outpatients' clinic	Seronegative for antimitochondrial antibodies; markers of hepatitis B/C virus infection; other possible causes of liver disease	Not specified
Larussa	2012	Investigate TPMT genotype and phenotype status in southern Italian IBD patients	Patients with Crohn's or UC	Blood transfusions within previous 3 months	Not specified
Lennard	2012	Investigate phenotype-genotype TPMT concordance in children with ALL	Patients diagnosed with ALL in time frame specified, at treatment centres in the UK and Ireland	Red cell transfusions within previous 2 months	Jan 1997 - Jun 2002
Liang	2013	"Investigate the relationship between TPMT enzymatic activity and genetic variation in TPMT with AZA clinical efficacy, especially in prevention of	Heart transplant recipients at Mayo Clinic; treated with AZA	Patients receiving MMF; dual organ recipients; cardiac amyloidosis patients; patients treated with drugs known to compete with AZA	First six months following cardiac transplant

Author	Year	Primary objective	Inclusion criteria	Exclusion criteria	Study period
		Rejection and safety in HTX recipients"		For metabolism	
Loennechen	2001	Identify TPMT mutant alleles in a Saami population as a basis of developing genotyping tests for prediction of TPMT activity	Patients >18 years old	Not specified	3 year period (1996 - 1999)
Ma	2006	Investigate the relationship between the TPMT gene polymorphisms and its enzymatic activity	Healthy blood donors; cord blood; patients with leukemia	Not specified	ALL patients: Dec 1995 - Jan 2002 (7 years); Cord blood samples: Jun 2001 - Aug 2001
Marinaki	2003	Establish frequencies of genetic modifiers of TPMT activity in Asian residents of the UK	Patients originating from India and Pakistan attending an IBD clinic	Not specified	(1) April 1999 - October 1999; (2) not specified
Milek	2006	Determine the frequency of clinically significant, low-activity TPMT alleles	Unrelated healthy volunteers	Not specified	Not specified
Oselin	2006	Develop and validate a HPLC method with UV detection to determine TPMT activity in human erythrocytes using 6-MP as a substrate	Volunteers; Estonian	Not specified	Not specified
Schaeffeler	2004	Sensitivity, specificity, PPV, and NPV for TPMT genotyping	No regular drug use with the exception of oral contraceptives and/or vitamins.	Not specified	Not specified
Schwab	2002	Whether AZA-related serious side effects can be explained by TPMT polymorphism using both pheno and geno typing	Patients with IBD from Department of Gastroenterology at University Hospital Tubingen;	Not specified	Not specified

Author	Year	Primary objective	Inclusion criteria	Exclusion criteria	Study period
			On AZA therapy at present or previously		
Serpe	2009	Elucidate the impact of genotype, age, gender on TPMT phenotype by comparing the activity of the enzyme among infants, children, adolescents, and adults	Healthy, unrelated, Italian-Caucasian adults; newborn, Italian-Caucasian babies, children, or adolescents	History of acute, chronic, genetic disease and/or medications, including anemia	Not specified
Spire-Vayron de la Moureyre	1998	Describe and demonstrate the usefulness of a new SSCP procedure to assay simultaneously for known mutations within TPMT, and to detect new ones	Selected from previously phenotyped individuals; healthy volunteers or patients	Not specified	Unclear
Spire-Vayron de la Moureyre	1998	Overall mutational spectrum of TPMT gene	Unrelated, European, volunteers or patients starting AZA therapy	Not specified	Not specified
von Ahsen	2005	Analyze AZA tolerance in relation to ITPA and TPMT mutation status and TPMT activity	>18 years; active Crohn's disease; prednisone treatment >300mg during the last 4 weeks or a relapse within 6 months after steroid pulse therapy	Not specified	2000 - 2002
Wennerstrand	2013	Investigate the fluctuation in TPMT enzyme activity from the time of diagnosis until after the end of maintenance treatment	Children starting their treatment per NOPHO ALL-2000 study protocol	Not specified	Not specified
Winter	2007	To determine if screening for TPMT status predicts side-effects to AZA in patients with IBD	Patients with IBD; no history of treatment with thiopurine drugs	Not specified	Feb 2003 - Jul 2006

Author	Year	Primary objective	Inclusion criteria	Exclusion criteria	Study period
Wusk	2004	Phenotype-genotype comparison of the TPMT enzyme; develop a new screening strategy for patients prior to taking thiopurine drugs	Unrelated healthy volunteers; patients with IBD	Not specified	Not specified
Xin	2009	Whether AZA-related serious side effects can be explained by the TPMT polymorphism using both phenotype and genotype tests in adult patients with renal transplantation on AZA therapy	Renal transplant recipients treated with AZA presently or previously	Not specified	Not specified
Yates	1997	Establish frequencies of the genetic modifiers of TPMT activity in an Asian population resident in the UK	Volunteer blood donors; children with ALL being treated or referred for evaluation	Not specified	Adults: 2 month period; Children: Not specified
Zhang	2007	Phenotype-genotype comparison of the TPMT enzyme and develop a new screening strategy for patients prior to taking thiopurine drugs	Patients with chronic renal failure; no blood transfusion within 1 month prior to study	Diabetic; neoplasm; active inflammations	Not specified

Abbreviations: AiH (autoimmune hepatitis); ALL (acute lymphocytic leukemia); AZA (azathioprine); HPLC (high performance liquid chromatography); IBD (inflammatory bowel disease); ITPA (inosine triphosphatase); NPV (negative predictive value); PPV (positive predictive value); QA (quality assurance); RBC (red blood cell); SSCP (single strand conformational polymorphism); TPMT (thiopurine s-methyltransferase); UC (ulcerative colitis); UK (United Kingdom); UV (ultraviolet); 6-MP (6-mercaptopurine)

Table 13. Design characteristics of high quality genotype-genotype studies

Author	Year	Primary objective	Inclusion criteria	Exclusion criteria	Time period
Chowdury	2007	Study compared three methods of genotyping - conventional vs microchip RFLP, and used TaqMan as the "gold standard". Also tested new steps in AS-PCR-CE and Portable microchip CE, but these were not tested against the others.	Patients with IBD; undergoing thiopurine immunosuppression	Not specified	Not specified
Kim	2013	Develop and validate a new AS-PCR for TPMT genotyping	Not specified	Not specified	Not specified
Lu	2005	Test feasibility of genotyping using APEX	Patients with b-thalassemia and random selection of patients for TPMT screening (healthy blood donors and children with ALL)	Not specified	Not specified
Ma	2003	To confirm and study the Chinese TPMT gene polymorphism; to compare and discuss the methodology for SNP tests; to find the best way and most suitable way to test the TPMT polymorphisms	ALL patients who were admitted inpatients by the Hematology Department of Beijing Children Hospital	Not specified	Dec 1995- Jan 2002
Roman	2012	To validate a TPMT genotyping method by comparing it with a conventional PCR approach	Adult white patients form the Hospital Universitario de la Princesa 9SPain) for whom genotyping was requested	Not specified	2006-2010
Schaeffeler	2008	Establishment and application of a novel assay, called iPLEX, for detection of all functional relevant 22 TPMT allelic variants	Healthy unrelated volunteers; Korean, Ghanians.	Not specified	Not specified

Abbreviations: ALL (acute lymphocytic leukemia); APEX (arrayed primer extension technology); AS-PCR (allele-specific polymerase chain reaction); CE (capillary electrophoresis), PCR (polymerase chain reaction); TPMT (thiopurine s-methyltransferase)

3.3.2 Sample characteristics

Sample characteristics for the high quality phenotype-genotype studies are presented in Table 14. Sample sizes ranged from 35 to 7195 individuals. Only four studies reported samples as pediatric [31, 35, 38, 45] and 11 studies did not specify the age of the sample population [16, 29, 30, 33, 34, 36, 40, 43, 68, 74, 86]. The remaining studies reported either adult or a mix of adult and pediatric sample populations. Race was not always specified [15, 33-36, 39, 43, 44, 67], but several high quality studies identified their sample population as Caucasian, Scandinavian, or from the UK [9, 28, 37, 38, 45, 66, 68, 70, 72, 73], European [29, 31, 74], German [16, 70], Chinese [32, 47], and Tunisian, Indian, Estonian, Slovenian, Italian, and Spanish [30, 40-42, 46, 68].

Sample characteristics for the high quality genotype-genotype studies are presented in Table 15. Sample sizes ranged from 80 to 630. Two studies included a mix of children and adults [77, 84], while one included adults only [81], and the remaining three did not specify the age group [78, 82, 83]. None of the studies specified the mean age of their subjects. One study was composed of solely IBD patients [78] and one did not specify the disease group of subjects [82].The remaining studies had a variety of subjects including ALL, otherwise healthy blood donors, and unspecified patients who were undergoing thiopurine treatment or who had TPMT testing requested. One study was in Chinese subjects [77], two were in white subjects [81, 82], and the other three did not specify a race or ethnicity [78, 83, 84]. Only two studies specified that subjects were unrelated [77, 82].

Table 14. Sample characteristics of high quality phenotype-genotype studies

Author	Year	Number included	Age group	Average age	Disease group	Ethnicity	Relation
Ben Salah	2013	88	Not specified	Not specified	Other	Tunisian	Not specified
Fakhoury	2007	118	Pediatric	6.12	Acute lymphoblastic leukemia	European	Not specified
Fangbin	2012	499	Adult	31.8	Inflammatory bowel disease	Chinese Han nationality; lived in Henan Province, Peoples Republic of China	Unrelated
Ford	2006	402	Not specified	Not specified	Not specified	Not specified	Not specified
Ford	2009	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified
Ganiere-Monteil	2004	468	Mix of adult and pediatric	40.9 (adult); 5.7 (child); 40 wGA (neonates)	Otherwise healthy	Caucasian	Not specified
Gazouli	2012	108	Pediatric	11.5	Inflammatory bowel disease	Not specified	Not specified
Hindorf	2012	7195	Not specified	Not specified	Inflammatory bowel disease	Not specified	Not specified
Jorquera	2012	200	Adult	Not specified	Otherwise healthy	Spanish, Chilean	Unrelated

Author	Year	Number included	Age group	Average age	Disease group	Ethnicity	Relation
Langley	2002	53	Mix of adult and pediatric	Not specified	Other	Not specified	Not specified
Larussa	2012	51	Adult	Not specified	Inflammatory bowel disease	Caucasian, Italian	Not specified
Lennard	2012	1117	Pediatric	Not specified	Acute lymphoblastic leukemia	UK and Ireland (English, Irish)	Not specified
Liang	2013	93	Adult	Not specified	Organ transplant	Not specified	Not specified
Loennechen	2001	260	Adult	Not specified	Patients admitted to a cardiology centre	Caucasian, Saami	Unrelated
Ma	2006	630	Mix of adult and pediatric	Not specified	Acute lymphoblastic leukemia	Chinese	Unrelated
Marinaki	2003	85	Not specified	Not specified	Inflammatory bowel disease	(1) Originating from India and Pakistan; (2) Caucasian	Not specified
Milek	2006	95	Not specified	Not specified	Otherwise healthy	Slovenian	Unrelated
Oselin	2006	99	Adult	32	Otherwise healthy	Estonian	Not specified
Schaeffeler	2004	1214	Adult	Not specified	Otherwise healthy	Caucasian, German	Unrelated
Schwab	2002	93	Adult	40-42 (depending on	Inflammatory bowel disease	Caucasian	Not specified

Author	Year	Number included	Age group	Average age	Disease group	Ethnicity	Relation
				group)			
Serpe	2009	943	Mix of adult and pediatric	0-68 (range)	Otherwise healthy	Italian-Caucasian	Unrelated
Spire-Vayron de la Moureyre	1998	35	Not specified	Not specified	Otherwise healthy	European	Not specified
Spire-Vayron de la Moureyre	1998	191	Not specified	Not specified	Not specified	European	Unrelated
von Ahsen	2005	71	Adult	Not specified	Inflammatory bowel disease	Caucasian	Not specified
Wennerstrand	2013	53	Pediatric	Not specified	Acute lymphoblastic leukemia	Scandinavian (Norway, Sweden, Finland)	Not specified
Winter	2007	130	Not specified	45	Inflammatory bowel disease	Not specified	Not specified
Wusk	2004	240	Not specified	Not specified	Inflammatory bowel disease	German	Unrelated
Xin	2009	150	Not specified	Not specified	Organ transplant	Not specified	Not specified
Yates	1997	48	Mix of adult and pediatric	Not specified	Acute lymphoblastic leukemia	Caucasian	Unrelated
Zhang	2007	278	Adult	Not specified	Other	Not specified	Unrelated

Table 15. Sample characteristics of high quality genotype-genotype studies

Author	Year	Number included	Age group	Average age	Disease group	Ethnicity	Relation
Chowdury	2007	80	Not specified	Not specified	IBD	Not specified	Not specified
Kim	2013	244	Not specified	Not specified	Requiring aza or mercaptopurine	Not specified	Not specified
Lu	2005	200	Children and adult	Not specified	B-thalassemia + pt selected for TPMT screening, also healthy volunteers		Not specified
Ma	2003	630	Mix of adult and pediatric	Not specified	ALL+ healthy blood donors, cord blood	Chinese	Unrelated
Roman	2012	111	Adult	Not specified	for whom TPMT genotyping was requested' - GE, derm, rheu, neph, inter med, hemato	White	Not specified
Schaeffeler	2008	586	Not specified	Not specified	Not specified	German (white)	Unrelated

Abbreviations: ALL (acute lymphocytic leukemia); IBD (inflammatory bowel disease); TPMT (thiopurine s-methyltransferase)

3.4 Laboratory test methods

The 30 high quality phenotype-genotype studies identified several different genotyping and phenotyping laboratory approaches. Table 16 reports the laboratory methods used for the test comparators for each high quality study. Studies often used more than one form of genotyping depending on the polymorphism of interest.

The high quality studies used only four different methods of phenotyping (Table 16), including RC method, high pressure liquid chromatography (HPLC), competitive micro-well immunoassay and mass spectrometry. Choice of phenotype method did not appear to be related to the study population (disease or age group). Most phenotype tests used either red blood cells (RBC) or RBC lysate, or white blood cells. In six studies, the sample was not as clearly specified, and was reported as 'whole blood', 'blood sample', or it was 'not specified'.

Similarly, it was not clear that the unit of analysis was different based on the phenotype laboratory method. For example, studies using the RC method reported TPMT activity in U/mL RBC [35, 36], and nmol/(ml RBC h) [73]. In addition the cutpoints for high pressure liquid chromatography differed between studies [16, 30, 41, 47, 66]. Some of these cutpoints were derived from previously cited laboratory work while others were calculated using receiver operating characteristics (ROC) after the testing was complete [32, 33, 40, 41]. Additional details of genotyping and phenotyping laboratory methods are provided below.

3.4.1 Genotyping

With regard to genotyping, studies employed similar DNA amplification methods, with 80% (24/30) of studies using PCR, 26% (8/30) using AS-PCR and 6% (2/30) using PCR-SSCP. Methods such as DHPLC [70, 72], ARMS [33, 34], pyrosequencing [36, 45], and TaqMan SNP genotyping [39, 40] were reported. Direct sequencing (n=3) [16, 29, 74] and RFLP (n=17) [9, 15, 28, 30, 32, 35, 37, 38, 40-44, 46, 47, 67, 68] were also reported. Only one study did not specify their genotyping method [73].

There were nine different methods of genotype testing reported (Tables 16 and 17). These included pyrosequencing (2/30), RFLP (includes restriction mapping, restriction analysis, or restriction digestion) (17/30), DHPLC (2/30), AS-PCR (8/30), direct sequencing (3/30), PCR-

SSCP (2/30) ARMS (2/30), PCR (24/30) and TaqMan methods (2/30). Twenty-six studies reported more than one method of genotyping, and one study did not report any method at all [73]. These methods will not be explained in detail here.

Table 17 presents the genotype laboratory methods for the six high quality genotype-genotype studies. Genotype-genotype test comparisons varied, and included several genotyping methods including RFLP [77, 78, 84], arrayed primer extension technology (APEX) [84], ARMS-PCR [84], AS-PCR [77, 78, 83], DHPLC [77, 82], LightSNiP [81], MALDI-TOF MS [82], PCR [77, 81, 83], SNaPshot sequencing [77], and TaqMan SNP genotyping [78]. Microchip RFLP and AS-PCR technologies were investigated in one study [78], and two studies referred to ‘sequencing’ as the genotyping method [77, 81].

All but one genotype-genotype study investigated at least TPMT*2 and TPMT*3, the most common polymorphisms [77, 78, 81-83]. One study investigated nearly all of the known TPMT polymorphisms, ten in total [82].

Genotype analysis can generally be categorized into three phases: DNA amplification, detection, and interpretation of the information gained. The most common method of amplification is PCR, however, there are also variations such as AS-PCR, ARMS, PCR-SSCP (single strand conformation/chain polymorphism), and whole genome amplification.

Table 16. Genotype and phenotype laboratory methods for high quality phenotype-genotype studies

Author	Year	Amplification/Genotype Method	Phenotype Method
Ben Salah	2013	PCR; AS-PCR; RFLP	HPLC
Fakhoury	2007	PCR; AS-PCR	HPLC
Fangbin	2012	PCR; RFLP	Not specified
Ford	2006	ARMS; AS-PCR; PCR	HPLC
Ford	2009	ARMS; AS-PCR; PCR	HPLC
Ganiere-Monteil	2004	PCR; AS-PCR	HPLC
Gazouli	2012	PCR; RFLP	RC
Hindorf	2012	Pyrosequencing	RC
Jorquera	2012	PCR; RFLP	HPLC
Langley	2002	PCR; RFLP	RC
Larussa	2012	PCR; RFLP	Competitive micro-well immunoassay
Lennard	2012	PCR; RFLP	HPLC
Liang	2013	PCR; TaqMan	Not specified
Loennechen	2001	PCR; AS-PCR; RFLP	RC
Ma	2006	PCR; RFLP	HPLC
Marinaki	2003	PCR; RFLP	RC
Milek	2006	PCR; RFLP, TaqMan	HPLC
Oselin	2006	PCR; RFLP	HPLC
Schaeffeler	2004	PCR; DHPLC	RC
Schwab	2002	DHPLC	Not specified

Author	Year	Amplification/Genotype Method	Phenotype Method
Serpe	2009	AS-PCR; PCR; RFLP	Not specified
Spire-Vayron de la Moureyre	1998	PCR-SSCP; Direct sequencing	RC
Spire-Vayron de la Moureyre	1998	PCR-SSCP; Direct sequencing	RC
von Ahsen	2005	Not specified	RC
Wennerstrand	2013	Pyrosequencing	RC
Winter	2007	PCR; RFLP	Mass spectrometry
Wusk	2004	PCR; sequencing	HPLC
Xin	2009	AS-PCR; PCR; RFLP	HPLC
Yates	1997	PCR; RFLP	RC
Zhang	2007	PCR; RFLP	HPLC

Abbreviations: ARMS (multiplex amplifications refractory mutation); AS-PCR (allele-specific polymerase chain reaction); DHPLC (denaturing high performance liquid chromatography); HPLC (high performance liquid chromatography); PCR (polymerase chain reaction); RC (radiochemical method); RFLP (restriction fragment length polymorphism)

Table 17. Genotype laboratory methods for high quality genotype-genotype studies

Author	Year	Index Method	Reference Method(s)	Polymorphisms Tested
Chowdury	2007	Microchip RFLP	Conventional RFLP and AS-PCR; Integrated Microchip PCR and AS-PCR; TaqMan SNP Genotyping	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Kim	2013	AS-PCR	PCR	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Lu	2005	APEX	ARMS-PCR; PCR-RFLP	TPMT*3b, TPMT*3c, TPMT*6
Ma	2003	PCR + DHPLC	PCR + RFLP; PCR + SNaPshot Sequencing with direct DNA sequencing; AS-PCR	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3d
Roman	2012	LightSNiP	Traditional PCR and Sangers Sequencing	TPMT*2, TPMT*3b, TPMT*3c
Schaeffeler	2008	MALDI-TOF MS	DHPLC	TPMT*2, TPMT*3a, TPMT*3c, TPMT*9, TPMT*11, TPMT*16, TPMT*17, TPMT*18, TPMT*20, TPMT*22

Abbreviations: AS-PCR (allele-specific polymerase chain reaction); DHPLC (denaturing high performance liquid chromatography); MALDI-TOF MS (matrix-assisted laser desorption ionization-time of flight mass spectrometry); PCR (polymerase chain reaction); RFLP (restriction fragment length polymorphism); SNP (single nucleotide polymorphism); TMPT (thiopurine s-methyltransferase)

3.4.1.1 Amplification of DNA

PCR works by amplifying and replicating a segment of DNA so that it is a more manageable segment to use. A component called Taq (or Hot Taq Star or U Taq) is a component used in PCR methodology to facilitate the bonding of DNA segments. 'Taq' refers to a highly specific thermal stable enzyme used to reduce the amplification of non-specific DNA. Applied at the appropriate time and temperature, it can improve the specificity of DNA amplification. PCR is generally used in all methods of genotyping, with whole genome amplification used for microarray as the primary exception.

AS-PCR uses primers which sit on a specific base in order to specify DNA replication. This particular base is at the end of the primer, and critical to identifying the appropriate allele. This is also known as 'matching' or 'binding'. AS-PCR will only amplify the DNA segment if there is a match with the specific allele of interest and the primer. As such, AS-PCR both amplifies and detects specific alleles in the same step. This method is beneficial in multiplexing and is potentially less costly. Multiplex ARMS is similar to AS-PCR in that it binds to and amplifies specific bases.

3.4.1.2 Choice of detection method

The next step in genotype testing is choosing a detection method. Detection methods can include RFLP (which includes restriction mapping, restriction analysis, restriction digest), TaqMan SNP genotyping, mass spectrometry, DHPLC, and pyrosequencing. RFLP uses bacteria to digest ('cut') the DNA strand at a certain point, thereby recognizing a sequence, which is followed by electrophoresis. RFLP may increase variability as there are multiple steps involved in maintaining the sample, necessitating more quality control steps and checks. This method is more difficult to use in high-throughput scenarios [87]. This detection method also requires more manipulation than other methods, and takes more time than a method such as TaqMan SNP genotyping. There may also be issues with storage, and technicians need to be careful with ensuring quality checks. RFLP can also be adapted to fluorescence. It is one of the older methods, and is less common in the present day laboratory for genotype testing [87].

TaqMan SNP genotyping uses two sets of reactions (two tests) as SNPs at different locations (or as many tests as there are SNPs). It takes place after amplification and uses two probes, i.e. a specific sequence of bases, each of which is fluoresced with two types of fluorescence (one

blue, one green). Fluorescence can be used in multiple approaches (such as TaqMan or RFLP) and is a method of assigning colour to visualize a target sequence or base. A ‘quencher’ (base at the end of the sequence) must match, in which case the dye hybridizes and shows colour if it matches. The Hospital for Sick Children, where this review was conducted, uses TaqMan technology, and routinely tests for TPMT*2 and TPMT*3 (personal communication, Tara Paton).

Mass spectrometry is also used for genotyping, although it is more commonly used for phenotype testing. Mass spectrometry measures the mass of each SNP. This can be particularly beneficial when looking at multiple variants at the same time. DHPLC has a separating mechanism, and is used in phenotype testing as well. Pyrosequencing is more of a quantitative sequencing, showing peak height, and is used for methylation and giving quantitative information, such as how much of an allele or how many bases are there, versus categorical data such as the presence or absence of an allele (as TaqMan provides).

3.4.1.3 Detection of genetic variants

Through these multiple methods of genotype testing, researchers have identified 21 TPMT specific polymorphisms that are associated with deficient or low TPMT activity, although with the ongoing advancement of genotype technology, more rare variants are likely to be discovered. TPMT*2 and TPMT*3 are the most common polymorphisms and make up more than 90 percent of polymorphisms. Table 18 outlines the polymorphisms investigated in each of the included studies. All but two studies comparing phenotype and genotype testing found in this review tested for (at least) TPMT*2 and TPMT*3. The outlying studies did not test for TPMT*2, only TPMT*3 [38, 67]. Although this increased the bias of these studies, the rest of the study characteristics were considered of high quality according to the QUADAS-2 and these studies were retained in the review.

Table 18. Genotype test characteristics and polymorphisms tested in phenotype-genotype studies

Author	Year	Population	Sample Source	Amplification/ Genotype Method	Polymorphisms Tested
Ben Salah	2013	Other	Erythrocytes (red blood cells)	PCR; AS-PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Fakhoury	2007	European	Blood samples including erythrocytes (red blood cells) and leucocytes (white blood cells), not otherwise state	PCR; AS-PCR	TPMT*2, TPMT*3a, TPMT*3c
Fangbin	2012	Chinese	Erythrocytes (red blood cells)	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Ford	2006	Not specified	Whole blood and RBC lysates (for comparison of methods)	ARMS; AS-PCR; PCR	TPMT*2, TPMT*3
Ford	2009	Not specified	Whole blood	ARMS; AS-PCR; PCR	TPMT*2, TPMT*3a, TPMT*3c
Ganiere-Monteil	2004	Caucasian	Erythrocytes (red blood cells)	PCR; AS-PCR	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Gazouli	2012	Not specified	Blood	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Hindorf	2012	Not specified	Whole blood	Pyrosequencing	TPMT*2, TPMT*3a, TPMT*3c; those with phenotype under 9.0 were further investigated on exons 3-10.
Jorquera	2012	Other	Blood	PCR; RFLP	TPMT*1, TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Langley	2002	Not specified	NR	PCR; RFLP	TPMT*3a, TPMT*3b, TPMT*3c
Larussa	2012	Caucasian	Lymphocytes	PCR; RFLP	TPMT*2, TPMT*3b, TPMT*3c
Lennard	2012	Other	NR	PCR; RFLP	TPMT*3a, TPMT*3b, TPMT*3c
Liang	2013	Not specified	Myocardial tissue or from blood samples	PCR; TaqMan	TPMT*2, TPMT*3a, TPMT*3c

Author	Year	Population	Sample Source	Amplification/ Genotype Method	Polymorphisms Tested
Loennechen	2001	Caucasian	Blood	PCR; AS-PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*6
Ma	2006	Chinese	Erythrocytes (red blood cells)	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3c
Marinaki	2003	Caucasian	Blood	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3c
Milek	2006	Other	Blood	PCR; RFLP, TaqMan	TPMT*2, TPMT*3b, TPMT*3c
Oselin	2006	Other	Whole blood	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D, TPMT*8
Schaeffeler	2004	Caucasian	Leukocytes (white blood cells)	PCR; DHPLC	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D
Schwab	2002	Caucasian	Leukocytes (white blood cells)	DHPLC	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D
Serpe	2009	Other	Whole blood	AS-PCR; PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Spire-Vayron de la Moureyre	1998	European	Leukocytes (white blood cells)	PCR-SSCP; Direct sequencing	TPMT*1, TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*1S, TPMT*1A, TPMT*7, TPMT *3d
Spire-Vayron de la Moureyre	1998	European	DNA from leukocytes (white blood cells)	PCR-SSCP; Direct sequencing	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D, TPMT*4, TPMT*5, TPMT*6, TPMT*7
von Ahsen	2005	Caucasian	Described elsewhere, reference provided	Not specified	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Wennerstrand	2013	Other	Whole blood	Pyrosequencing	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c, TPMT*3D

Author	Year	Population	Sample Source	Amplification/ Genotype Method	Polymorphisms Tested
Winter	2007		Whole blood	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Wusk	2004	German	Blood	PCR; Sequencing	TPMT*2, TPMT*3b, TPMT*3c
Xin	2009	Not specified	Leukocytes (white blood cells)	AS-PCR; PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c
Yates	1997	Caucasian	Leukocytes (white blood cells)	PCR; RFLP	TPMT*1, TPMT*2, TPMT*3a, TPMT*3c
Zhang	2007	Not specified	Leukocytes (white blood cells)	PCR; RFLP	TPMT*2, TPMT*3a, TPMT*3b, TPMT*3c

Abbreviations: ARMS (multiplex amplification refractory mutation); AS-PCR (allele-specific polymerase chain reaction); DHPLC (denaturing high performance liquid chromatography); NR (not reported); PCR (polymerase chain reaction); RFLP (restriction fragment length polymorphism); SSCP (single strand conformation polymorphism); TMPT (thiopurine s-methyltransferase)

Table 19. Phenotype laboratory methods and cutpoints for high quality studies

Author	Year	Disease	Phenotype Method	Sample Source	Substrate	Cutpoints	Unit
Ben Salah	2013	Crohn's Disease	HPLC	Erythrocyte lysate (red blood cell lysate)	6-MP	Low (not specified), intermediate (5-10), high (>10),	nmol 6-MMP/h/ml pRBC
Fakhoury	2007	Acute lymphoblastic leukemia	HPLC	Erythrocytes (red blood cells)	6-MP	Intermediate (<11.8); deficient estimated from graph as approximately 6	U/mL pRBCs
Fangbin	2012	IBD	Not specified	Erythrocytes (red blood cells)	6-MP	Optimal cutoff calculated by ROC: intermediate (<4.75) (heterozygous carrier).	U/mL RBC
Ford	2006	Not specified	HPLC	Whole blood and erythrocytes (red blood cells) (for methods comparison)	6-TG	Researchers calculated own cutpoint for low/intermediate; unclear whether they calculated it for high	nmol 6MTG/gHb/h
Ford	2009	Not specified	HPLC	Whole blood	Not specified	Deficient (<5); low (6-34); normal (35-79); high (>=80)	nmol 6MTG/gHb/h
Ganiere-Monteil	2004	Otherwise healthy	HPLC	Erythrocyte lysate (red blood cell lysate)	6-MP	Post-hoc suggestion of phenotype cut-off (13.5) between wild-type and heterozygous genotype.	U/mL pRBC

Author	Year	Disease	Phenotype Method	Sample Source	Substrate	Cutpoints	Unit
Gazouli	2012	IBD	RC	Erythrocyte lysate (red blood cell lysate)	Not specified	Low (<5.5), intermediate (5.6-15.5); Normal-high (>15.6)	U/mL RBC
Hindorf	2012	IBD	RC	Erythrocytes (red blood cells)	6-MP	Low (<2.5); high (>9.0)	U/mL pRBC
Jorquera	2012	Otherwise healthy	HPLC	Blood sample	6-TG	Deficient (</=5); low (6-24); normal (25-55); high (>/=56)	nmol/gHb/h
Langley	2002	Autoimmune liver disease	RC	Erythrocyte lysate (red blood cell lysate)	6-MP	Deficient (<5.0); intermediate (5-13.7); high (>13.7)	U/ml
Larussa	2012	IBD	Competitive micro-well immunoassay	Erythrocytes (red blood cells)	6-MP	Very low (</=5.5); intermediate (5.6-15.5); normal to hi (>/=15.6)	U/gHb
Lennard	2012	Acute lymphoblastic leukemia	HPLC	Not specified	Not specified	Between intermediate and high - varied cutpoints at 9.5, 10.5, 11.5	Units/mL pRBC
Liang	2013	Organ transplant	Not specified	Erythrocytes (red blood cells)	6-MP	Low (<6.3); intermediate (6.3-15.0); normal (15.1-26.4)	U/ml RBC

Author	Year	Disease	Phenotype Method	Sample Source	Substrate	Cutpoints	Unit
Loennechen	2001	Patients admitted to a cardiology centre	RC	Erythrocytes (red blood cells)	Not specified	Deficient (<5); heterozygous intermediate (5-9.5); wild-type (>9.5)	U/mL pRBC
Ma	2006	Acute lymphoblastic leukemia	HPLC	Erythrocytes (red blood cells)	6-MP	12	U
Marinaki	2003	IBD and dermatology patients	RC	Erythrocytes (red blood cells)	6-MP	Low (<2.5); intermediate (2.5-8); normal (8-15)	nmol 6-MMP/h/ml RBC
Milek	2006	Otherwise healthy	HPLC	Erythrocytes (red blood cells)	6-MP	Calculated using ROC analysis; Low<5.8 assumed based on previous study as no LO found in this study so unable to calculate own cutpoint; High >9.82	pmol 6-MMP/ 10^7 RBC/h
Oselin	2006	Otherwise healthy	HPLC	Whole blood	6-MP	Researchers calculated own cutpoint using ROC; between high-intermediate (58.8)	ng/ml/h
Schaeffeler	2004	Otherwise healthy	RC	Erythrocytes (red blood cells)	6-TG	Low (<9), intermediate (9-22); high (22-50); very high (51-65)	nmol

Author	Year	Disease	Phenotype Method	Sample Source	Substrate	Cutpoints	Unit
Schwab	2002	IBD	Not specified	Erythrocytes (red blood cells)	6-TG	High (>24); low (<3)	nmol 6-MTG/ gHb /h
Serpe	2009	Otherwise healthy	Not specified	Erythrocyte lysate (red blood cell lysate)	6-MP	"arbitrary cutpoints" low (<8.0); intermediate (<19.4); normal (<37.0); high (>37.0)	U/gHb; nmol 6-MMP/h
Spire-Vayron de la Moureyre	1998	Otherwise healthy	RC	Erythrocyte lysate (red blood cell lysate)	Not specified	Deficient (<5 U/ml); intermediate (5-13.7), high (>13.7),	U/ml RBC
Spire-Vayron de la Moureyre	1998	Not specified	RC	Erythrocyte lysate (red blood cell lysate)	6-MP	Low (<5); intermediate (5-13.7); high (>13.7)	U/mL RBC
von Ahsen	2005	IBD	RC	Erythrocyte lysate (red blood cell lysate)	Other	Low (<10)	nmol/(mL RBC/h)
Wennerstrand	2013	Acute lymphoblastic leukemia	RC	Blood sample	6-MP	Low vs intermediate (2.5); high vs intermediate (9.0)	U/mL pRBC
Winter	2007	IBD or UC	Mass spectrometry	Whole blood	Not specified	Low (<10), intermediate (10-25); normal (26-50); high (>50)	pmol/h/mg Hb
Wusk	2004	IBD	HPLC	Erythrocyte lysate (red blood cell lysate)	6-MP	Heterozygous carrier (45.5)	nmol MTG/gHb/h

Author	Year	Disease	Phenotype Method	Sample Source	Substrate	Cutpoints	Unit
Xin	2009	Organ transplant	HPLC	Erythrocytes (red blood cells)	6-TG	Very low (<3); intermediate (3-24); normal (24-50); high (>50U)	U
Yates	1997	Acute lymphoblastic leukemia	RC	Erythrocytes (red blood cells)	Not specified	Deficient (<5.0); heterozygous (5-10); homozygous wild-type (>10)	U/ml pRBC
Zhang	2007	Chronic renal failure	HPLC	Erythrocytes (red blood cells)	6-MP	Calculated by ROC	nmol/ml pRBC

Abbreviations: gHb (gram of hemoglobin); h (hour); HPLC (high performance liquid chromatography); IBD (inflammatory bowel disease); ml (millilitre); MTG (methylthioguanine); ng (nanogram); nmol (nanomole); pRBC (packed red blood cells); pmol (picomole); RBC (red blood cell); RC (Radiochemical method); ROC (receiver operating characteristic); U (unit); UC (ulcerative colitis); 6-MMP (6-methyl-mercaptopurine); 6-MP (6-mercaptopurine); 6-TG (6-thioguanine)

3.4.2 Phenotype tests

Phenotype test methods included RC method (11/30), HPLC (13/30), competitive micro-well immunoassay (1/30), and mass spectrometry (1/30), with four studies unclear about the method they used for phenotype testing. Phenotype laboratory methods and cutpoints for high quality studies are presented in Table 19. In addition to RC assay, HPLC, immunoassay, and mass spectrometry, there were also variations such as tandem mass spectrometry and modifications to the traditional RC assay. The RC assay can be conducted using either 6-mercaptopurine (6-MP) or 6-tioguanine (6-TG) as substrates, both of which are thiopurines. The addition of fluorescence can improve specificity of detection [87].

Measurement units for reporting enzyme activity varied across studies. Most commonly, enzyme activity was measured per milliliter of pRBCs (U/mL pRBCs). Alternatively, it was measured as nanomoles of 6-methylthioguanine (MTG) per gram hemoglobin per hour (nmol 6-MTG/gHb/h), international units per ml (IU/mL), picomole (pmol) 6-MP (6-MMP)/10^7 RBC/h, pmol/minute/milligram protein, or U/gHb. Variation in units made direct comparison of enzyme activity cutpoints across studies difficult.

In general, TPMT activity was classified by authors as low, intermediate, or high activity. However, as discussed previously, terminology and classification of activity levels was inconsistent, with some studies using ’deficient’ where some authors used ’low’, some studies adding a category of ’very high’, and some studies using ‘normal’ in place of ’high’. Table 14, 16, and 19 describe the characteristics of the phenotype tests.

The choice of cutpoint was generally cited from previous research and literature, although some researchers calculated their own cutpoints after sample collection and analysis. Typically this was in the form of a ROC analysis [32, 33, 40, 41, 44]. The conventional classification system developed by Weinshilboum et al. [8] in the 1980’s classifies phenotype activity as deficient (<5 U/ml RBC), intermediate (5-10 U/ml pRBC), and normal (>10 U/ml pRBC). This classification was used in three studies [9, 28, 30].

It was not clear whether cutpoints varied by any particular study characteristic or population. For example, the cutpoint between intermediate and high enzyme activity for populations of ALL

patients varied from 9 to 12 U/mL pRBCs, while the cutpoint between intermediate and low varied between 2.5 to 6 U/mL pRBCs. For patients with IBD, the cutpoint between intermediate and high enzyme activity varied between 8 and 45.5 nmol 6-MTG/gHb/h, or 4.75 and 15.5 U/mL RBC. Also for IBD patients, the cutpoint between low and intermediate varied between 2.5 and 5.6 U/mL RBC. In contrast to these values, one study reported a cutpoint of 25 between intermediate and high enzyme activity and a cutpoint of 10 between low and intermediate, however, the unit of this test was specified as picomoles [15]. Further, some studies did not specify the unit of measure.

3.5 Diagnostic test performance characteristics

Diagnostic test performance characteristics, such as sensitivity, specificity, NPV, PPV, and concordance, were infrequently explicitly reported in studies comparing two testing approaches. In cases where authors did report test performance characteristics, a concordance rate was commonly reported. Seven studies reported sensitivity and specificity explicitly and additional nine reported NPV, PPV, or concordance. These are reported in Table 20.

Using raw data from all 30 high quality phenotype-genotype studies, the sensitivity, specificity, NPV, PPV and concordance were calculated with genotyping as the index test and phenotype testing set as the reference standard. Table 21 presents test performance characteristics for genotyping when ‘deficient’ was defined as the absence of TPMT activity (suggesting the presence of a homozygous mutation). With data that were available from 15 studies, calculated sensitivity of genotyping ranged from 0.0 to 100.0% and with data that were available from 26 studies, specificity ranged from 97.8 to 100.0%.

Fifteen studies provided data sufficient to calculate both sensitivity and specificity for detection of a homozygous mutation [9, 15, 29-31, 33, 35-37, 42, 66, 67, 70, 72, 74]. Due to the absence of homozygous deficient patients (cell count of zero), it was not possible to calculate either sensitivity or specificity for the remaining studies. Ten of the 15 studies with sufficient data had 100.0% for both values [9, 29-31, 33, 42, 66, 70, 72, 74]. The other five studies only investigated TPMT*2 and TPMT*3, although half of those studies with 100.0% calculated values also were limited to these polymorphisms [30, 31, 33, 42, 66]. The two studies with a sensitivity of 0.0% were conducted in samples of 130 (total persons with positive test for low enzyme

activity = 1, total persons with negative test for low enzyme activity = 129) [15] and 53 (total persons with positive test for low enzyme activity = 1, total persons with negative test for low TPMT activity - 52) [67] patients, respectively. Most studies with calculated sensitivity and specificity of 100.0% generally had large sample sizes (n = 88 to1214) (total persons with positive test for low enzyme activity ranged from 1 to 7, and total persons with negative test for low enzyme activity ranged from 34 to 1207). The largest study [36] had a sensitivity of 86.0% and tested for all polymorphisms.

Four of five studies with imperfect sensitivity and specificity did not specify the race of the population studied [15, 35, 36, 67]. Among the six studies where polymorphisms beyond the common TPMT*2 and TPMT*3 were examined [29, 36, 45, 70, 72, 74], five were conducted in European populations; - the sixth did not specify the ethnicity or race of its sample population [36].

Table 22 presents test performance characteristics for genotyping when ‘deficient’ was defined as absent to intermediate TPMT activity (suggesting the presence of a homozygous OR heterozygous mutation). This table was easier to populate compared to the previous table (Table 21) as the ability to detect mutations increased with the added consideration of the heterozygous mutation, which is more commonly found. Therefore, there were fewer missing values. Calculated sensitivity ranged from 13.4 to 100.0% and specificity ranged from 90.9 to 100.0%. Twenty-five studies provided sufficient data to calculate both sensitivity and specificity. Of the 25 studies, only one study had perfect sensitivity and specificity of 100.0%, the only study conducted in a Tunisian population [30].

There was no clear trend indicating whether additional SNPs increased the sensitivity, Six of nine (67%) studies with >75% sensitivity tested only TPMT*2 and TPMT*3 whereas 12/16 (75%) of studies with <75% sensitivity tested only TPMT*2 and TPMT*3. One study with >75% sensitivity had a sample size of 35 (total persons with low + intermediate enzyme activity = 18, total persons with high enzyme activity = 17) [74], while the remaining eight ranged from 88 to 1214 (total persons with low + intermediate enzyme activity ranged from 5 to 954, total persons with high enzyme activity ranged from 17 to 6241) [9, 28, 30, 33, 46, 66, 70, 72].

Only four studies in the genotype-genotype comparison group reported test performance characteristics. Roman et al. [81] reported sensitivity, specificity, PPV, and NPV in their study. Schaefffeler et al. [70] , Lu et al. [84], and Anglicheau et al. [11] reported concordance (Table 23).

Table 20. Diagnostic test performance values as reported by authors

Author	Year	Amplification/ Genotype Method	Phenotype Method	Purpose of Test	Reported Sensitivity	Reported Specificity	Reported PPV	Reported NPV	Reported Concordance
Fangbin	2012	PCR; AS-PCR; RFLP	Not specified	To predict a heterozygous carrier	100	97.3	Not specified	Not specified	Not specified
Ford	2006	PCR; AS-PCR	HPLC	Phenotype versus genotype for normal TPMT and *1/*1	Not specified	Not specified	Not specified	Not specified	98.1
Ford	2009	PCR; RFLP	HPLC	def/Ho; low/He; norm/WT	Not specified	Not specified	Not specified	Not specified	mean: 95; 83; 68
Gazouli	2012	ARMS; AS-PCR; PCR	RC	Overall concordance between genotype and phenotype	Not specified	Not specified	Not specified	Not specified	88.2
Hindorf	2012	ARMS; AS-PCR; PCR	RC	Overall concordance between genotype and phenotype	Not specified	Not specified	Not specified	Not specified	95
Lennard	2012	PCR; AS-PCR	HPLC	To detect variant	78	97	Not specified	Not specified	92 overall
Liang	2013	PCR; RFLP	Not specified	Between He and INT	Not specified	Not specified	Not specified	Not specified	89
Milek	2006	Pyrosequencing	HPLC	Genotype to phenotype concordance rate	Not specified	Not specified	Not specified	Not specified	91.6
Oselin	2006	PCR; RFLP	HPLC	To predict wildtype and heterozygous individuals	89%	88%	Not specified	Not specified	Not specified

Author	Year	Amplification/ Genotype Method	Phenotype Method	Purpose of Test	Reported Sensitivity	Reported Specificity	Reported PPV	Reported NPV	Reported Concordance
Schaeffeler	2004	PCR; RFLP	RC	Concordance between heterozygosity and intermediate phenotype; wild-type and normal/high phenotype	Not specified	Not specified	Not specified	Not specified	89.2% (heterozygous and intermediate); 99.4% (wild-type and normal/high)
Serpe	2009	PCR; RFLP	Not specified	Overall concordance	Not specified	Not specified	Not specified	Not specified	71
Spire-Vayron de la Moureyre	1998	PCR; RFLP	RC	Agreement with deficient methylator	Not specified	Not specified	Not specified	Not specified	100
Winter	2007	PCR; TaqMan	Mass spectrometry	For genotyping, including all identified mutations to predict phenotype	90	99	94	99	Not specified
Wusk	2004	PCR; AS-PCR; RFLP	HPLC	To predict phenotype	100%	89%	42%	Not specified	Not specified
Yates	1997	PCR; RFLP	RC	To predict intermediate TPMT activity and heterozygous phenotype	95	100	Not specified	Not specified	Not specified
Zhang	2007	PCR; RFLP	HPLC	To predict carrier	100	95%	37%	Not specified	Not specified

Abbreviations: ARMS (multiplex amplification refractory mutation); AS-PCR (allele-specific polymerase chain reaction); HPLC (high performance liquid chromatography); PCR (polymerase chain reaction); RC (radiochemical method); RFLP (restriction fragment length polymorphism); TPMT (thiopurine s-methyltransferase)

Table 21. Diagnostic test performance values from 2x2 tables for presence of homozygous deficient genotype

Author	Year	Amplification/ Genotype Method	Phenotype Method	Calculated Sensitivity	Calculated Specificity	Calculated PPV	Calculated NPV
Ben Salah	2013	PCR; AS-PCR; RFLP	HPLC	100.0%	100.0%	100.0%	100.0%
Fakhoury	2007	PCR; AS-PCR	HPLC	100.0%	100.0%	100.0%	100.0%
Fangbin	2012	PCR; RFLP	Not specified	*	100.0%	*	100.0%
Ford	2006	ARMS; AS-PCR; PCR	HPLC	100.0%	100.0%	100.0%	100.0%
Ford	2009	ARMS; AS-PCR; PCR	HPLC	*	*	*	*
Ganiere-Monteil	2004	PCR; AS-PCR	HPLC	100.0%	100.0%	100.0%	100.0%
Gazouli	2012	PCR; RFLP	RC	33.3%	98.9%	85.7%	88.1%
Hindorf	2012	Pyrosequencing	RC	86.0%	100.0%	100.0%	99.9%
Jorquera	2012	PCR; RFLP	HPLC	*	100.0%	*	100.0%
Langley	2002	PCR; RFLP	RC	0.0%	100.0%	*	98.1%
Larussa	2012	PCR; RFLP	Competitive micro-well immunoassay	16.7%	97.8%	50.0%	89.8%
Lennard	2012	PCR; RFLP	HPLC	*	*	*	100.0%
Liang	2013	PCR; TaqMan	Not specified	*	100.0%	*	100.0%
Loennechen	2001	PCR; AS-PCR; RFLP	RC	*	100.0%	*	100.0%
Ma	2006	PCR; RFLP	HPLC	*	100.0%	*	100.0%
Marinaki	2003	PCR; RFLP	RC	*	100.0%	*	100.0%

Author	Year	Amplification/ Genotype Method	Phenotype Method	Calculated Sensitivity	Calculated Specificity	Calculated PPV	Calculated NPV
Milek	2006	PCR; RFLP, TaqMan	HPLC	*	100.0%	*	100.0%
Oselin	2006	PCR; RFLP	HPLC	*	*	*	*
Schaeffeler	2004	PCR; DHPLC	RC	100.0%	100.0%	100.0%	100.0%
Schwab	2002	DHPLC	Not specified	100.0%	100.0%	100.0%	100.0%
Serpe	2009	AS-PCR; PCR; RFLP	Not specified	100.0%	100.0%	100.0%	100.0%
Spire-Vayron de la Moureyre	1998	PCR-SSCP; Direct sequencing	RC	100.0%	100.0%	100.0%	100.0%
Spire-Vayron de la Moureyre	1998	PCR-SSCP; Direct sequencing	RC	100.0%	100.0%	100.0%	100.0%
von Ahsen	2005	Not specified	RC	*	100.0%	*	100.0%
Wennerstrand	2013	Pyrosequencing	RC	*	100.0%	*	100.0%
Winter	2007	PCR; RFLP	Mass spectrometry	0.0%	100.0%	*	99.2%
Wusk	2004	PCR; sequencing	HPLC	*	.	*	*
Xin	2009	AS-PCR; PCR; RFLP	HPLC	*	100.0%	*	100.0%
Yates	1997	PCR; RFLP	RC	100.0%	100.0%	100.0%	100.0%
Zhang	2007	PCR; RFLP	HPLC	*	100.0%	*	100.0%

Abbreviations: ARMS (multiplex amplification refractory mutation); AS-PCR (allele-specific polymerase chain reaction); HPLC (high performance liquid chromatography); PCR (polymerase chain reaction); RC (radiochemical method); RFLP (restriction fragment length polymorphism)

*Unable to calculate

Table 22. Diagnostic test performance from 2x2 tables for presence of homozygous or heterozygous deficient genotype

Author	Year	Amplification/ Genotype Method	Phenotype Method	Calculated Sensitivity	Calculated Specificity	Calculated PPV	Calculated NPV
Ben Salah	2013	PCR; AS-PCR; RFLP	HPLC	100.0%	100.0%	100.0%	100.0%
Fakhoury	2007	PCR; AS-PCR	HPLC	29.3%	97.5%	85.7%	72.6%
Fangbin	2012	PCR; RFLP	Not specified	38.5%	100.0%	100.0%	97.3%
Ford	2006	ARMS; AS-PCR; PCR	HPLC	80.6%	98.1%	80.6%	98.1%
Ford	2009	ARMS; AS-PCR; PCR	HPLC	*	*	*	*
Ganiere-Monteil	2004	PCR; AS-PCR	HPLC	92.7%	100.0%	100.0%	99.3%
Gazouli	2012	PCR; RFLP	RC	52.2%	100.0%	100.0%	73.8%
Hindorf	2012	Pyrosequencing	RC	69.5%	98.8%	89.5%	95.5%
Jorquera	2012	PCR; RFLP	HPLC	83.3%	99.5%	93.8%	98.4%
Langley	2002	PCR; RFLP	RC	66.7%	90.9%	60.0%	93.0%
Larussa	2012	PCR; RFLP	Competitive micro-well immunoassay	22.2%	97.0%	80.0%	69.6%
Lennard	2012	PCR; RFLP	HPLC	*	*	*	92.2%
Liang	2013	PCR; TaqMan	Not specified	60.0%	98.7%	90.0%	92.8%
Loennechen	2001	PCR; AS-PCR; RFLP	RC	95.8%	100.0%	100.0%	99.6%
Ma	2006	PCR; RFLP	HPLC	67.7%	99.8%	95.5%	98.4%
Marinaki	2003	PCR; RFLP	RC	55.6%	100.0%	100.0%	95.0%

Author	Year	Amplification/ Genotype Method	Phenotype Method	Calculated Sensitivity	Calculated Specificity	Calculated PPV	Calculated NPV
Milek	2006	PCR; RFLP, TaqMan	HPLC	50.0%	97.6%	75.0%	93.1%
Oselin	2006	PCR; RFLP	HPLC	*.	*	*	*
Schaeffeler	2004	PCR; DHPLC	RC	86.8%	99.4%	94.9%	98.4%
Schwab	2002	DHPLC	Not specified	100.0%	96.6%	62.5%	100.0%
Serpe	2009	AS-PCR; PCR; RFLP	Not specified	13.4%	98.3%	78.8%	70.3%
Spire-Vayron de la Moureyre	1998	PCR-SSCP; Direct sequencing	RC	83.3%	94.1%	93.8%	84.2%
Spire-Vayron de la Moureyre	1998	PCR-SSCP; Direct sequencing	RC	54.5%	94.3%	66.7%	90.9%
von Ahsen	2005	Not specified	RC	*	100.0%	*	75.8%
Wennerstrand	2013	Pyrosequencing	RC	17.4%	100.0%	100.0%	59.6%
Winter	2007	PCR; RFLP	Mass spectrometry	64.7%	100.0%	100.0%	95.0%
Wusk	2004	PCR; sequencing	HPLC	*	*	*	*
Xin	2009	AS-PCR; PCR; RFLP	HPLC	29.2%	100.0%	100.0%	88.1%
Yates	1997	PCR; RFLP	RC	96.3%	100.0%	100.0%	95.5%
Zhang	2007	PCR; RFLP	HPLC	36.8%	100.0%	100.0%	95.6%

Abbreviations: ARMS (multiplex amplification refractory mutation); AS-PCR (allele-specific polymerase chain reaction); DHPLC (denaturing high performance liquid chromatography); HPLC (high performance liquid chromatography); PCR (polymerase chain reaction); RC (radiochemical method); RFLP (restriction fragment length polymorphism)

*Unable to calculate

Table 23. Reported diagnostic test performance for high quality genotype-genotype comparisons

Author	Year	Index Test	Reference Test(s)	Purpose of Test	Reported Sensitivity	Reported Specificity	Reported PPV	Reported NPV	Reported Concordance
Chowdury	2007	Microchip RFLP	Conventional RFLP and AS-PCR; Integrated Microchip PCR and AS-PCR; TaqMan SNP Genotyping	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified
Kim	2013	AS-PCR	PCR	Not specified	Not specified	Not specified	Not specified	Not specified	"in agreement"
Lu	2005	APEX	ARMS-PCR or PCR-RFLP	Not specified	Not specified	Not specified	Not specified	Not specified	100.0
Ma	2003	PCR + DHPLC	PCR + RFLP; PCR + SNaPshot Sequencing with direct DNA sequencing; AS-PCR	Not specified	Not specified	Not specified	Not specified	Not specified	Not specified
Roman	2012	LightSNiP	Traditional PCR and Sangers sequencing	Conventional method vs LightSNiP	100.0	100.0	97.0	Not specified	Not specified
Schaeffeler	2008	MALDI-TOF MS	DHPLC	Concordance with previous study genotype results	Not specified	Not specified	Not specified	Not specified	100.0

Abbreviations: APEX (arrayed primer extension technology); ARMS (multiplex amplification refractory mutation); AS-PCR (allele-specific polymerase chain reaction); DHPLC (denaturing high performance liquid chromatography); MALDI-TOF MS (matrix-assisted laser desorption ionization-time of flight mass spectrometry); PCR (polymerase chain reaction); RFLP (restriction fragment length polymorphism); SNP (single nucleotide polymorphism);

4 DISCUSSION

4.1 Systematic review and quality appraisal

The choice of technologies available for the diagnosis of TPMT deficiency is varied. This review revealed a diverse and large body of literature assessing both phenotype and genotype technologies for TPMT testing across several disease states. Literature exists comparing phenotype and genotype technologies, as well as comparing different laboratory methodologies within each technology (genotype-genotype testing, and phenotype-phenotype testing).

This detailed systematic review revealed that the inclusion in the search strategies of diagnostic test terms such as ‘sensitivity’, ‘specificity’, ‘diagnostic error’, or ‘accuracy’, in combination with the other primary search concepts (TPMT, genotype or phenotype test, and thiopurine drugs), resulted in exclusion of several relevant papers. Conversely, it appeared that relevant studies were classified in the citation databases under ‘TPMT’ rather than the specific thiopurine drugs, as the inclusion of ‘TPMT’ in combination with ‘any thiopurine drug’ increased the inclusion of appropriate papers.

Many of the studies screened aimed to examine the TPMT test result (TPMT status) in relation to ADE rates. ADEs of interest included myelosuppression (including leukopenia, neutropenia, thrombocytopenia), pancreatitis, and hepatotoxicity. However ADEs can result from factors others than TPMT status. In addition, a health care practitioner’s tolerance of risk of ADE depends on the disease in question and treatment goals. Therefore, an analysis of the relationship between ADE and TPMT status was not included in the present review.

There were also several foreign language studies that emerged from the searching, particularly in German and Chinese that were included in the review. The exclusion of foreign language papers would have otherwise resulted in the exclusion of relevant studies, and is an important consideration when conducting reviews on this subject. Also, for the purpose of this review, there was no need to limit studies of TPMT status to certain disease populations or age groups. A total of 66 studies that met inclusion criteria were retrieved that were published over the period of 1996 to 2014.

The review revealed that there are limitations to both genotype testing and phenotype testing, with neither test accepted as a 'gold standard' for identifying TPMT deficiency. The results from this review demonstrate the multitude of inquiries into which method is more accurate, with increasing focus on genotype methods in recent years.

A recent systematic review of guidelines for TPMT testing demonstrated wide variation across clinical sub-specialties regarding the choice of TPMT testing method. There was also variation in recommended adjustments in dosing of thiopurines in the event of a positive finding, as well as the quality of the guideline development process [25]. The present review may serve to provide clarity to assist decisions in the choice of testing method.

The quality appraisal revealed that the quality of the studies was varied. Inadequate reporting of information regarding index tests, reference tests, recruitment methods, and study populations were the primary reasons for exclusion of studies due to quality. Although the QUADAS tool includes an applicability assessment, the focus of the research question resulted in few studies being rejected based on this element.

There was a paucity of reporting by authors of diagnostic precision (sensitivity, specificity, NPV, and PPV values), indicating a need for guidance on reporting of test performance characteristics for diagnostic technologies. This review resulted in a total of 30 high quality studies comparing phenotype and genotype technologies, with sufficient data to assess the diagnostic accuracy of these technologies. There were an additional six high quality genotype-genotype studies.

4.2 TPMT test performance characteristics

Many studies did not report performance characteristics in terms of sensitivity, specificity, PPV, or NPV. In the cases where performance characteristics were reported, it was rare for 95% confidence intervals around the estimates to be reported.

Rather than use reported information, the sensitivity and specificity of the genotype test was calculated for the high quality studies based on data extracted from individual studies. The low prevalence of deficient TPMT activity (homozygous mutations) in the population made it challenging to acquire a sufficient or appropriate population such that diagnostic test accuracy

could be calculated for many studies. Of the high quality studies, only three studies [35, 36, 70] had more than two subjects categorized as positive for deficient TPMT activity and homozygous mutations.

This report found that a number of studies selectively conducted a genotype test only for those subjects who had low TPMT enzyme activity, introducing bias into the calculation of test performance characteristics. This choice may be related to the comparatively high cost of genotyping, however, the serial testing design inflated genotype test sensitivity.

Among the high quality studies, the number of polymorphisms included in genotype tests ranged from two to nine, with most studies including TPMT*2 and TPMT*3, which are the most common genetic variants in persons with deficient TPMT activity [9]. As the number of polymorphisms tested for increased, the sensitivity of the test was expected to increase, and with the exception of one study this trend was generally shown. Studies testing for up to three polymorphisms had calculated sensitivity between 22.2% and 69.5% (n=11), whereas studies testing five or more polymorphisms had a calculated sensitivity of 54.5% to 100.0% (n=6). The exception was a study investigating five polymorphisms [45] which had a sample size of 53 (total persons with positive test for low enzyme activity = 23, total persons with negative test for low enzyme activity = 28) and a calculated sensitivity of 17.4%. The highest quality studies assessed genotype tests that tested for TPMT*1 (1S & 1A), TPMT*2, TPMT*3(3A, 3B, 3C & 3D), TPMT*4, TPMT*5, TPMT*6, TPMT*7, and TPMT*8.

With regard to measurement of enzymatic activity for the phenotype test, limited consistency in cutpoints between low, intermediate and high activity categories was observed. Authors frequently used a ROC analysis to determine the cutpoint for their study population. In addition, measurement units for enzyme activity were variable, making the comparability of cutpoints difficult.

Using a cutpoint that defined ‘deficient’ as the absence of enzyme activity/presence of a homozygous mutation, the calculated sensitivity ranged from 0.0% to 100.0% and the calculated specificity ranged from 97.8% to 100.0%. Among the fifteen studies for which both sensitivity and specificity could be calculated, 10 demonstrated perfect (100%) sensitivity and specificity. The inference of perfect values may be misleading, however. Due to the low prevalence of

homozygous mutations (0.3%), it is possible that the sample sizes of the studies were too small for a stable rate of detection of this rare mutation, hindering an accurate calculation of sensitivity and specificity.

Using a cutpoint that defined deficient as the low to intermediate enzyme activity/presence of heterozygous or homozygous mutation, the calculated sensitivity ranged from 17.4% to 100.0% and the calculated specificity ranged from 90.9% to 100.0%. Only one of twenty-five studies for which both sensitivity and specificity could be calculated displayed perfect (100%) sensitivity and specificity. Raising the cutpoint for the definition of 'deficient' activity to include the absence of activity (homozygous mutation) and intermediate activity (heterozygous mutation) enabled the detection of more positive cases, resulting in more stable determinations of sensitivity and specificity from the data provided.

The clinical utility of TPMT testing lies in its ability to distinguish patients with homozygous mutations (deficient TPMT activity) from other patients to know in whom thiopurines should be avoided. Only 15 studies included sufficient data to estimate sensitivity and specificity of genotyping for this purpose. It is also important to distinguish heterozygous patients (intermediate TPMT activity) from homozygous and from wild type patients to identify individuals who can receive thiopurines, but who require a reduced dose. It was evident from the review that distinguishing between these different patient groups was not the priority in many of these studies.

The variation in sensitivity and specificity observed in the present review may also be related to the disease context. In more severe and life-threatening diseases such as ALL, a higher risk of drug-related adverse events such as myelosuppression may be tolerated to maximize the therapeutic dose of the thiopurine. This would result in a preference for a higher threshold resulting in more false negatives (lower sensitivity) and fewer false positives (higher specificity). A different set of thresholds, and consequently values for sensitivity and specificity, may be preferred for chronic disease such as IBD and dermatological conditions.

4.3 Comparison to previous reviews

In a previous review conducted by TASK, 17 studies of the performance characteristics of phenotype or genotype testing strategies were identified [2], however, not all of these studies were found to be of high quality when appraised using the QUADAS-2 tool in this review. Nine studies included in the previous review were appraised as low quality and excluded from this review [10, 11, 13, 14, 57, 59, 61, 74, 85]. In the previous review, the genotype test performance characteristics, expressed in terms of sensitivity and specificity, ranged from 55 to 100% and from 94 to 100%, respectively. The sensitivity and specificity of the phenotype test ranged from 92 to 100% and 86 to 98%, respectively.

Poor reporting practices was a significant contributor to the exclusion of studies from the present review and was also found in the previous TPMT review which used a modified Critical Appraisal Skills Program tool [26]. In another systematic review of papers studying the relationship between genotype and drug-related myelosuppression, a quality appraisal of 67 studies using published guidelines designed to assess quality of pharmacogenetic studies) [88] did not detect any low quality studies [89]. In a review comparing phenotype and genotype diagnostic accuracy where the QUADAS-2 tool was used to appraise quality of studies, 37% of the studies reviewed were deemed low quality [90]. The range of quality appraisal tools, reporting practices, and judgements regarding high and low quality underscore the importance of addressing quality of reporting in diagnostic studies, as well optimal choice of quality appraisal tools for diagnostic studies. This issue will become more salient considering the increasing use of pharmacogenomics in health care. A genomics domain was deliberately added to the QUADAS-2 quality appraisal tool for the present review to address the risk of bias associated with studies assessing genomic diagnostic tests.

A comprehensive MA of 16 studies of TPMT test performance was performed by the US Agency for Healthcare Research and Quality (AHRQ) in 2010. The MA was performed to derive a pooled estimate of TPMT genotyping performance characteristics for identifying individuals with low or intermediate TPMT activity or with any mutation in the TPMT gene. In that review, pooled sensitivity for detecting homozygosity and heterozygosity was 70.7% (95% confidence interval 37.90 to 90.50) and pooled specificity was 99.9% (95% confidence interval 97.40 to 99.60). Sensitivity and specificity estimates from individual studies were statistically transformed to

make them more normally distributed before independent mean estimates were calculated [90]. However, that review did not address the correlation between sensitivity and specificity in performing the MA. A further limitation of the AHRQ analysis was that it only considered TPMT testing for IBD patients and omitted adults or children with ALL. In addition, the AHRQ analysis clearly stated that it assumed all cut-points for labeling results as positive or negative were the same across studies. The variation in cut-points observed in the present review suggests that an assumption of equivalent cut-points may have introduced bias into the AHRQ pooled estimates [90].

4.4 Laboratory Methods

Consideration of pre-analytical components is important to the success of any diagnostic test, as the risk of error in the laboratory is highest during this phase [90]. Both phenotype and genotype tests contain laboratory and operator steps which could affect the possibility of error.

The choice of primer and other laboratory conditions can affect genotyping and potentially sensitivity and specificity. Choosing an appropriate primer requires knowledge of the genotype method to be used and population that is being sampled. Several factors, such as temperature, reaction timing, proximity of the primer to the target sequence, reagent choice, and operator technique can affect how well the primer binds, or whether it binds at all, to the target sequence [87, 91, 92]. In the event that a poor choice of primer or poor methodology is employed, a false negative result may occur if the primer is unable to bind and identify the sequence.

Phenotype testing has known confounders which can result in false positives. Recent blood transfusions can give a falsely high indication of the patient's true TPMT activity status, and it is recommended that there is a 120 day window between blood transfusion and phenotypic TPMT activity measurement. Similarly, certain medications are known to interact with thiopurines, such as allopurinol and 5-aminosalicylates, and affect TPMT activity measurement, and co-administration should be avoided to reduce risk of ADE [93] .

Genotyping offers a solution to the variability of TPMT phenotype activity measurement and potential misclassification due to confounding variables. Graham [92] suggests that selectively genotyping of patients whose phenotype tests indicate low enzyme activity may be a solution to

the confounding issues of phenotype testing, for patients who may be at highest risk of an ADE. Genotyping would then provide confirmation of TPMT activity status. Again, the issue of choice of polymorphisms in the genotype must be considered.

4.5 Study strengths and limitations

The literature search strategy was designed to be comprehensive, capturing all possible citation databases as well as numerous sources for grey literature. In addition, translations of foreign articles were sought so that translation bias would not be present. Nevertheless, it is possible that some articles were missed.

The bulk of filtering of abstracts and titles, as well as reviewing full studies for eligibility and appraising the quality of studies, were performed by a single reviewer (LR). A second reviewer verified the eligibility of uncertain studies and independently reviewed and appraised a 5% random subset as a quality control measure. The present review would be enhanced had two independent reviewers been available for all filtering, review and appraisal tasks.

Choosing the QUADAS-2 allowed the assessment to be tailored to the specific research question. The QUADAS tool is currently recommended by the Cochrane Diagnostic Test Accuracy Working Group, a worldwide leader in systematic review and quality appraisal [94]. One disadvantage of the QUADAS tool is that it can be described as a summary tool that was not designed to distinguish between low and high quality. As such, it required the reviewers to develop criteria regarding what constitutes a low quality paper. The creation of these criteria may be a source of variability between different study groups. The addition of a genomics domain to the QUADAS-2 appraisal tool significantly improved the ability to use this diagnostic accuracy test quality appraisal tool to assess bias pertaining to genomic testing. Although there are intrinsic elements of freedom in the QUADAS-2 tool, with open-ended descriptions and the use of flow diagrams created by the user, the criteria for decisions of bias and applicability were quite stringent. The creation of a genomics domain can be useful for future quality appraisals of studies assessing genetic or genomic diagnostic tests.

In all possible cases sensitivity, specificity, NPV, and PPV were calculated. However, the calculations of sensitivity and specificity were hampered by the absence of reported cell count

data in many studies. In addition, low cell counts may have contributed to unstable estimates. A minimum sample size calculation or required cell count were not recorded as components of the quality appraisal. For studies that assess diagnostic tests of rare variants, a minimum sample size or cell count for calculation of sensitivity and specificity should be established and included as a quality criterion.

In the absence of a gold standard, the present review set the reference test as the phenotype test. This is the older test and test results are subject to confounding from blood transfusions as well as potential drug interactions [93], and known non-perfect sensitivity and specificity [26]. The range of polymorphisms included in the genotype test would also affect its sensitivity and specificity, thus both approaches have limitations.

4.6 Implications

There is a growing use of personalized medicine applications such as pharmacogenomics in clinical diagnostics and clinical decision-making for selection of drug treatment and dose. The increasing complexity of genomic technologies for pharmacogenomics is associated with greater cost. Routine testing for all possible polymorphisms is more costly and unlikely to be feasible for health care institutions. Although current tests may become less costly in the future, there may also be mutations that have not yet been identified with current methods. Next generation sequencing including whole exome and whole genome sequencing are expected to provide greater yield of genetic variants related to disease as well as drug metabolizing enzymes [95], but use of these technologies may not be cost-effective for all applications and requires further evaluation. Consideration also needs to be given to oversight and regulation, the applicability of pharmacogenetic discoveries in ethnically diverse populations and special populations (children, elderly), and to the anticipated shift from diagnostic pharmacogenetic testing to screening. Lack of agreement on the clinical implementation of pharmacogenetic testing for TPMT persists [25].

Clinical and institutional decision-makers require high quality evidence of clinical validity and clinical utility of TPMT genotyping technologies to ensure appropriate and consistent use in patient populations who would benefit from this testing. This review showed a lack of a clear assessment of test performance characteristics for the purpose of identifying patients with

deficient enzyme activity (for drug avoidance) or identifying patients with intermediate enzyme activity (for dose reduction).

Government decision-makers who decide on reimbursement for genotyping testing also require high quality evidence of cost-effectiveness. With growing demands on limited health care budgets, there is a need for methods to fully assess the social, legal, ethical as well as economic implications of TPMT genotyping. Health technology assessment plays an important role in generating this evidence and in promoting consistent reporting methods to facilitate the evaluation process.

4.7 Future research

Many studies did not report sufficient data to accurately calculate study performance metrics, despite being designed for that purpose. There is a need for consistent guidelines for reporting findings in order to evaluate the accuracy of diagnostic tests. This will be increasingly important as new technologies evolve, such as next generation sequencing. Likewise, it is important that these future studies sample subjects with homozygous mutations and deficient TPMT activity to better estimate sensitivity of diagnostic tests.

It is important to perform a MA of the data obtained in this systematic review. A MA can combine the available data to obtain joint overall estimates of sensitivity and specificity for both phenotype and genotype testing along with uncertainty ranges. Understanding the performance of both tests, as well as the uncertainty associated with those estimates, can help guide adoption decisions or decisions to perform more research. Recent meta-analytic techniques that address the lack of a gold standard can be used to address the major challenge of assessing the phenotype and genotype TPMT tests [96-99].

5 CONCLUSIONS

There is a growing use of personalized medicine applications such as pharmacogenomics in clinical diagnostics and clinical decision-making for selection of drug treatment and dose to avert serious adverse drug events. This review of the literature comparing phenotype testing and genotype testing for TPMT status demonstrates a broad base of evidence for these tests. The quality of the studies for assessing diagnostic test accuracy was mixed. The literature

displayed a profound lack of patients with low TPMT activity or homogeneous TPMT mutations, making estimates of sensitivity of the tests uncertain. Clinical and institutional decision-makers require high quality evidence of clinical validity and clinical utility of TPMT genotyping technologies to ensure appropriate and consistent use in patient populations who would benefit from this testing.

REFERENCES

1. Tantisira, K. and S.T. Weiss, *Overview of pharmacogenomics*, in *UpToDate*, D.S. Basow, Editor 2013, UpToDate: Waltham, MA.
2. Donnan, J.R., et al., *Systematic review of thiopurine methyltransferase genotype and enzymatic testing strategies.* Ther Drug Monit, 2011. **33**(2): p. 192-9.
3. MacDermott, R.P., *6-mercaptopurine (6-MP) metabolite monitoring and TPMT testing in the treatment of inflammatory bowel disease with 6-MP or azathioprine*, in *UpToDate*, D.S. Basow, Editor 2013, UpToDate: Waltham, MA.
4. Sahasranaman, S., D. Howard, and S. Roy, *Clinical pharmacology and pharmacogenetics of thiopurines.* Eur J Clin Pharmacol, 2008. **64**(8): p. 753-67.
5. Baker, G.R., et al., *The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada.* Canadian Medical Association Journal, 2004. **170**(11): p. 1678-1686.
6. Lathia, N., et al., *Evaluation of direct medical costs of hospitalization for febrile neutropenia.* Cancer, 2010. **116**(3): p. 742-8.
7. Weycker, D., et al., *Cost of neutropenic complications of chemotherapy.* Ann Oncol, 2008. **19**(3): p. 454-60.
8. Weinshilboum, R.M. and S.L. Sladek, *Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thiopurine methyltransferase activity.* Am J Hum Genet, 1980. **32**(5): p. 651-62.
9. Yates, C.R., et al., *Molecular diagnosis of thiopurine S-methyltransferase deficiency: genetic basis for azathioprine and mercaptopurine intolerance.* Annals of Internal Medicine, 1997. **126**(8): p. 608-14.
10. Alves, S., et al., *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): p. 165-74.
11. Anglicheau, D., et al., *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): p. 119-27.
12. Indjova, D., et al., *Phenotypic and genotypic analysis of thiopurine S-methyltransferase polymorphism in the Bulgarian population.* Therapeutic Drug Monitoring, 2003. **25**(5): p. 631-636.
13. Kham, S.K.Y., et al., *Thiopurine S-methyltransferase activity in three major Asian populations: a population-based study in Singapore.* European Journal of Clinical Pharmacology, 2008. **64**(4): p. 373-9.
14. Larovere, L.E., et al., *Genetic polymorphism of thiopurine S-methyltransferase in Argentina.* Annals of Clinical Biochemistry, 2003. **40**(4): p. 388-393.
15. Winter, J.W., 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): p. 1069-77.
16. Wusk, B., et al., *Thiopurine S-methyltransferase polymorphisms: efficient screening method for patients considering taking thiopurine drugs.* European Journal of Clinical Pharmacology, 2004. **60**(1): p. 5-10.

17. Stanulla, M., et al., *Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia.* JAMA : the journal of the American Medical Association, 2005. **293**(12): p. 1485-9.
18. Ujiie, S., et al., *Functional characterization of 23 allelic variants of thiopurine S-methyltransferase gene (TPMT*2 - *24).* Pharmacogenet Genomics, 2008. **18**(10): p. 887-93.
19. Jones, C.D., et al., *Thiopurine methyltransferase activity in a sample population of black subjects in Florida.* Clin Pharmacol Ther, 1993. **53**(3): p. 348-53.
20. Kham, S.K., et al., *Thiopurine methyltransferase polymorphisms in a multiracial asian population and children with acute lymphoblastic leukemia.* J Pediatr Hematol Oncol, 2002. **24**(5): p. 353-9.
21. Lee, E.J. and W. Kalow, *Thiopurine S-methyltransferase activity in a Chinese population.* Clin Pharmacol Ther, 1993. **54**(1): p. 28-33.
22. Park-Hah, J.O., et al., *Thiopurine methyltransferase activity in a Korean population sample of children.* Clin Pharmacol Ther, 1996. **60**(1): p. 68-74.
23. Kongkaew, C., P.R. Noyce, and D.M. Ashcroft, *Hospital admissions associated with adverse drug reactions: a systematic review of prospective observational studies.* The Annals of pharmacotherapy, 2008. **42**(7): p. 1017-25.
24. Taylor-Gjevre, R.M., et al., *Thiopurine methyltransferase screening before azathioprine prescription: a physician survey.* J Popul Ther Clin Pharmacol, 2013. **20**(1): p. e13-7.
25. HF Burnett, R.T., W Chandranipapongse, P Madadi, S Ito, and WJ Ungar, *Testing for thiopurine methyltransferase status for safe and effective thiopurine administration: a systematic review of clinical guidance documents.* The Pharmacogenomics Journal, 2014: p. 1-10.
26. Donnan, J.R., et al., *Health Technology Assessment of Thiopurine Methyltransferase Testing for Guiding 6-Mercaptopurine Doses in Pediatric Patients with Acute Lymphoblastic Leukemia*, 2010, Technology Assessment at SickKids (TASK).
27. Whiting, P.F., et al., *QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.* Ann Intern Med, 2011. **155**(8): p. 529-36.
28. Loennechen, T., et al., *Detection of one single mutation predicts thiopurine S-methyltransferase activity in a population of Saami in northern Norway.* Clin Pharmacol Ther, 2001. **70**(2): p. 183-188.
29. Spire-Vayron de la Moureyre, C., et al., *Genotypic and phenotypic analysis of the polymorphic thiopurine S-methyltransferase gene (TPMT) in a European population.* British Journal of Pharmacology, 1998. **125**(4): p. 879-87.
30. Ben Salah, L., et al., *Analysis of thiopurine S-methyltransferase phenotype-genotype in a Tunisian population with Crohn's disease.* European Journal of Drug Metabolism & Pharmacokinetics, 2013. **38**(4): p. 241-4.
31. Fakhoury, M., et al., *Should TPMT genotype and activity be used to monitor 6-mercaptopurine treatment in children with acute lymphoblastic leukaemia?* Journal of Clinical Pharmacy & Therapeutics, 2007. **32**(6): p. 633-9.
32. Fangbin, Z., 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): p. 695-701 10.1097/FTD.0b013e3182731925.
33. Ford, L., V. Graham, and J. Berg, *Whole-blood thiopurine S-methyltransferase activity with genotype concordance: a new, simplified phenotyping assay.* Annals of Clinical Biochemistry, 2006. **43**(Pt 5): p. 354-60.

34. Ford, L., P. Kampanis, and J. Berg, *Thiopurine S-methyltransferase genotype-phenotype concordance: used as a quality assurance tool to help control the phenotype assay.* Annals of Clinical Biochemistry, 2009. **46**(Pt 2): p. 152-4.
35. Gazouli, M., et al., *Thiopurine methyltransferase genotype and thiopurine S-methyltransferase activity in Greek children with inflammatory bowel disease.* Annals of Gastroenterology, 2012. **25**(3): p. 249-253.
36. Hindorf, U. and M.L. Appell, *Genotyping should be considered the primary choice for pre-treatment evaluation of thiopurine methyltransferase function.* Journal of Crohn's & colitis, 2012. **6**(6): p. 655-9.
37. Larussa, T., et al., *High prevalence of polymorphism and low activity of thiopurine methyltransferase in patients with inflammatory bowel disease.* European Journal of Internal Medicine, 2012. **23**(3): p. 273-7.
38. Lennard, L., et al., *Thiopurine methyltransferase genotype-phenotype discordance and thiopurine active metabolite formation in childhood acute lymphoblastic leukaemia.* British Journal of Clinical Pharmacology, 2013. **76**(1): p. 125-36.
39. Liang, J.J., et al., *TPMT genetic variants are associated with increased rejection with azathioprine use in heart transplantation.* Pharmacogenetics and Genomics, 2013. **23**(12): p. 658-665.
40. Milek, M., et al., *Thiopurine S-methyltransferase pharmacogenetics: genotype to phenotype correlation in the Slovenian population.* Pharmacology, 2006. **77**(3): p. 105-14.
41. Oselin, K., et al., *Determination of thiopurine S-methyltransferase (TPMT) activity by comparing various normalization factors: reference values for Estonian population using HPLC-UV assay.* Journal of Chromatography B: Analytical Technologies in the Biomedical & Life Sciences, 2006. **834**(1-2): p. 77-83.
42. Serpe, L., et al., *Thiopurine S-methyltransferase pharmacogenetics in a large-scale healthy Italian-Caucasian population: differences in enzyme activity.* Pharmacogenomics, 2009. **10**(11): p. 1753-65.
43. Xin, H.-W., et al., *Relationships between thiopurine S-methyltransferase polymorphism and azathioprine-related adverse drug reactions in Chinese renal transplant recipients.* European Journal of Clinical Pharmacology, 2009. **65**(3): p. 249-55.
44. Zhang, L.-R., et al., *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): p. 45-51.
45. Wennerstrand, P., et al., *Methotrexate binds to recombinant thiopurine S-methyltransferase and inhibits enzyme activity after high-dose infusions in childhood leukaemia.* European Journal of Clinical Pharmacology, 2013. **69**(9): p. 1641-1649.
46. Jorquera, A., et al., *[Phenotype and genotype of thiopurine methyltransferase in Chilean individuals].* Revista Medica de Chile, 2012. **140**(7): p. 889-95.
47. Ma, X.L., et al., *Relationships between thiopurine methyltransferase gene polymorphisms and its enzymatic activity. [Chinese].* Zhonghua zhong liu za zhi [Chinese journal of oncology], 2006. **28**(6): p. 456-459.
48. Ansari, A., et al., *Thiopurine methyltransferase activity and the use of azathioprine in inflammatory bowel disease.* Alimentary Pharmacology & Therapeutics, 2002. **16**(10): p. 1743-50.
49. Arenas, M., et al., *Genetic variation in the MTHFR gene influences thiopurine methyltransferase activity.* Clinical Chemistry, 2005. **51**(12): p. 2371-4.

50. Barlow, N.L., V. Graham, and J.D. Berg, *Expressing thiopurine S-methyltransferase activity as units per litre of whole-blood overcomes misleading high results in patients with anaemia.* Annals of Clinical Biochemistry, 2010. **47**(Pt 5): p. 408-14.
51. Ebbesen, M.S., et al., *Incorporation of 6-thioguanine nucleotides into DNA during maintenance therapy of childhood acute lymphoblastic leukemia-the influence of thiopurine methyltransferase genotypes.* Journal of Clinical Pharmacology, 2013. **53**(6): p. 670-4.
52. Evans, W.E., et al., *Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine.* Journal of Clinical Oncology, 2001. **19**(8): p. 2293-301.
53. Ferucci, E.D., et al., *Azathioprine metabolite measurements are not useful in following treatment of autoimmune hepatitis in Alaska Native and other non-Caucasian people.* Canadian Journal of Gastroenterology, 2011. **25**(1): p. 21-7.
54. Gu, L.-j., et al., *[Significance of TPMT activity and TGNs level detection for individualizing 6-mercaptopurine chemotherapy].* Chung-Hua Hsueh Yeh Hsueh Tsa Chih: Chinese Journal of Hematology, 2003. **24**(1): p. 18-21.
55. Haglund, S., et al., *Pyrosequencing of TPMT alleles in a general Swedish population and in patients with inflammatory bowel disease.[Erratum appears in Clin Chem. 2004 Apr;50(4):788].* Clinical Chemistry, 2004. **50**(2): p. 288-95.
56. Heckmann, J.M., et al., *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): p. 71-80.
57. Hon, Y.Y., et al., *Polymorphism of the thiopurine S-methyltransferase gene in African-Americans.* Human Molecular Genetics, 1999. **8**(2): p. 371-6.
58. Kasirer, Y., et al., *Thiopurine S-methyltransferase (TPMT) Activity Is Better Determined by Biochemical Assay Versus Genotyping in the Jewish Population.* Dig Dis Sci, 2014.
59. Reis, M., A. Santoro, and G. Suarez-Kurtz, *Thiopurine methyltransferase phenotypes and genotypes in Brazilians.* Pharmacogenetics, 2003. **13**(6): p. 371-3.
60. Relling, M.V., et al., *Prognostic importance of 6-mercaptopurine dose intensity in acute lymphoblastic leukemia.* Blood, 1999. **93**(9): p. 2817-2823.
61. Rossi, A.M., et al., *Genotype-phenotype correlation for thiopurine S-methyltransferase in healthy Italian subjects.* European Journal of Clinical Pharmacology, 2001. **57**(1): p. 51-4.
62. Schmiegelow, K., et al., *Thiopurine methyltransferase activity is related to the risk of relapse of childhood acute lymphoblastic leukemia: results from the NOPHO ALL-92 study.* Leukemia, 2009. **23**(3): p. 557-64.
63. Sies, C., et al., *Measurement of thiopurine methyl transferase activity guides dose-initiation and prevents toxicity from azathioprine.* New Zealand Medical Journal, 1210. **118**(1210).
64. Tamm, R., et al., *Thiopurine S-methyltransferase (TPMT) pharmacogenetics: three new mutations and haplotype analysis in the Estonian population.* Clinical Chemistry & Laboratory Medicine, 2008. **46**(7): p. 974-9.
65. el-Azhary, R.A., et al., *Thioguanine nucleotides and thiopurine methyltransferase in immunobullous diseases: optimal levels as adjunctive tools for azathioprine monitoring.* Archives of Dermatology, 2009. **145**(6): p. 644-52.
66. Ganiere-Monteil, C., et al., *Phenotype and genotype for thiopurine methyltransferase activity in the French Caucasian population: impact of age.* European Journal of Clinical Pharmacology, 2004. **60**(2): p. 89-96.

67. Langley, P.G., et al., *Thiopurine methyltransferase phenotype and genotype in relation to azathioprine therapy in autoimmune hepatitis.* Journal of Hepatology, 2002. **37**(4): p. 441-7.
68. Marinaki, A.M., et al., *Genetic determinants of the thiopurine methyltransferase intermediate activity phenotype in British Asians and Caucasians.* Pharmacogenetics, 2003. **13**(2): p. 97-105.
69. Relling, M.V., et al., *Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus.* Journal of the National Cancer Institute, 1999. **91**(23): p. 2001-8.
70. Schaeffeler, E., 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): p. 407-17.
71. Schmiegelow, K., et al., *Methotrexate/6-mercaptopurine maintenance therapy influences the risk of a second malignant neoplasm after childhood acute lymphoblastic leukemia: results from the NOPHO ALL-92 study.* Blood, 2009. **113**(24): p. 6077-84.
72. Schwab, M., et al., *Azathioprine therapy and adverse drug reactions in patients with inflammatory bowel disease: impact of thiopurine S-methyltransferase polymorphism.* Pharmacogenetics, 2002. **12**(6): p. 429-36.
73. von Ahsen, N., 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): p. 2282-8.
74. Spire-Vayron de la Moureyre, C., 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): p. 177-85.
75. Lennard, L. and H.J. Singleton, *High-performance liquid chromatographic assay of human red blood cell thiopurine methyltransferase activity.* Journal of Chromatography B: Biomedical Applications, 1994. **661**(1): p. 25-33.
76. Lu, Y., et al., *Genotyping of eight polymorphic genes encoding drug-metabolizing enzymes and transporters using a customized oligonucleotide array.* Analytical Biochemistry, 2007. **360**(1): p. 105-113.
77. Ma, X.-L., et al., *[Exploration of methodology for assay of single nucleotide polymorphism in thiopurine methyltransferase gene].* Zhongguo Shi Yan Xue Ye Xue Za Zhi, 2003. **11**(5): p. 458-63.
78. Chowdhury, J., et al., *Microfluidic platform for single nucleotide polymorphism genotyping of the thiopurine S-methyltransferase gene to evaluate risk for adverse drug events.* Journal of Molecular Diagnostics, 2007. **9**(4): p. 521-9.
79. Kim, J.H., et al., *Influences of thiopurine methyltransferase genotype and activity on thiopurine-induced leukopenia in Korean patients with inflammatory bowel disease: a retrospective cohort study.* Journal of Clinical Gastroenterology, 2010. **44**(10): p. e242-8.
80. Osaki, R., et al., *Accuracy of genotyping using the TaqMan PCR assay for single nucleotide polymorphisms responsible for thiopurine sensitivity in Japanese patients with inflammatory bowel disease.* Experimental and Therapeutic Medicine, 2011. **2**(5): p. 783-786.
81. Roman, M., et al., *Validation of a genotyping method for analysis of TPMT polymorphisms.* Clinical Therapeutics, 2012. **34**(4): p. 878-84.
82. Schaeffeler, E., et al., *Highly multiplexed genotyping of thiopurine s-methyltransferase variants using MALD-TOF mass spectrometry: reliable genotyping in different ethnic groups.* Clinical Chemistry, 2008. **54**(10): p. 1637-47.

83. Kim, S., et al., *Validation of new allele-specific real-time PCR system for thiopurine methyltransferase genotyping in Korean population.* BioMed Research International, 2013. **2013**: p. 305704.
84. Lu, H.-F., Shih, M.-C. et al, *Molecular anlaysis of the thiopurine S-methyltransferase alleles in Bolivians and Tibetans.* Journal of Clinical Pharmacy and Therapeutics, 2005. **30**: p. 491-496.
85. Indjova, 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): p. 637-44.
86. Winter, J., et al., *Cost-effectiveness of thiopurine methyltransferase genotype screening in patients about to commence azathioprine therapy for treatment of inflammatory bowel disease.* Aliment Pharmacol Ther, 2004. **20**(6): p. 593-9.
87. Kucukkal, T.G., Yang Ye, Chapman Susan C, Cao Weiguo, Alexov Emil, *Computational and Experimental Approaches to Reveal the Effects of Single Nucleotide Polymorphisms with Respect to Disease Diagnostics.* International Journal of Molecular Sciences, 2014. **15**: p. 9670-9717.
88. Jorgensen, A.L.W., Paula R, *Methodological quality of pharmacogenetic studies: issus of concern.* Statistics in Medicine, 2008. **27**: p. 6547-6569.
89. Higgs, J.E., et al., *Are patients with intermediate TPMT activity at increased risk of myelosuppression when taking thiopurine medications?* Pharmacogenomics, 2010. **11**(2): p. 177-88.
90. Booth, R.A., 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.*, 2010: Rockville, MD: Agency for Healthcare Research and Quality.
91. Shin, G.W., Chung, Boram, Jung, Gyu Yong, Jung, GyooYeol, *Multiplex ligase-based genotyping methods combined with CE.* Electrophoresis, 2014. **35**: p. 1004-1016.
92. Graham, V., *Thiopurine methyltransferase Phenotyping and Genotyping in Clinical Practice*, in *College of Medical and Dental Sciences*2009, The University of Birmingham: Birmingham. p. 167.
93. Sanderson, J., et al., *Thiopurine methyltransferase: should it be measured before commencing thiopurine drug therapy?* Annals of clinical biochemistry, 2004. **41**(Pt 4): p. 294-302.
94. Reitsma, J., Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ., *Chapter 9: Assessing methodological quality.*, in *Cochrane Handbook for Systematic Reviews of Diagnostic Accuracy Version 1.0.0*, J. Deeks, Bossuyt PM, Gatsonis C (editors), Editor 2009. p. 28.
95. Wright, C., et al, *Next steps in the sequence: The implications of whole genome sequencing for health in the UK*, 2011, PHG Foundation: Cambridge, UK.
96. Albert, P.S., *Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard.* Stat Med, 2009. **28**(5): p. 780-97.
97. Dendukuri, N., et al., *Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference.* Biometrics, 2012. **68**(4): p. 1285-93.
98. Hawkins, D.M., J.A. Garrett, and B. Stephenson, *Some issues in resolution of diagnostic tests using an imperfect gold standard.* Stat Med, 2001. **20**(13): p. 1987-2001.
99. Rutjes, A.W., et al., *Evaluation of diagnostic tests when there is no gold standard. A review of methods.* Health Technol Assess, 2007. **11**(50): p. iii, ix-51.