SYSTEMATIC REVIEW

Guidance for the Harmonisation and Improvement of Economic Evaluations of Personalised Medicine

Heleen Vellekoop[1] · Simone Huygens[1] · Matthijs Versteegh[1] · László Szilberhorn[2] · Tamás Zelei[2] · Balázs Nagy[2] · Rositsa Koleva-Kolarova[3] · Apostolos Tsiachristas[3] · Sarah Wordsworth[3] · Maureen Rutten-van Mölken[1,4] on behalf of the HEcoPerMed Consortium

Abstract

Objective The objective of this study was to develop guidance contributing to improved consistency and quality in economic evaluations of personalised medicine (PM), given current ambiguity about how to measure the value of PM as well as considerable variation in the methodology and reporting in economic evaluations of PM.

Methods A targeted literature review of methodological papers was performed for an overview of modelling challenges in PM. Expert interviews were held to discuss best modelling practice. A systematic literature review of economic evaluations of PM was conducted to gain insight into current modelling practice. The findings were synthesised and used to develop a set of draft recommendations. The draft recommendations were discussed at a stakeholder workshop and subsequently finalised.

Results Twenty-two methodological papers were identified. Some argued that the challenges in modelling PM can be addressed within existing methodological frameworks, others disagreed. Eighteen experts were interviewed. They believed large uncertainty to be a key concern. Out of 195 economic evaluations of PM identified, 56% addressed none of the identified modelling challenges. A set of 23 recommendations was developed. Eight recommendations focus on the modelling of test-treatment pathways. The use of non-randomised controlled trial data is discouraged but several recommendations are provided in case randomised controlled trial data are unavailable. The parameterisation of structural uncertainty is recommended. Other recommendations consider perspective and discounting; premature survival data; additional value elements; patient and clinician compliance; and managed entry agreements.

Conclusions This study provides a comprehensive list of recommendations to modellers of PM and to evaluators and reviewers of PM models.

1 Introduction

Personalised medicine (PM) aims to better stratify patients to enable more targeted healthcare. Personalised medicine, often used interchangeably with related terms such as precision medicine, stratified medicine and individualised medicine, has the potential to offer cost savings (e.g. due to therapies being prescribed only to those likely to benefit) and improved health outcomes (e.g. due to dose adjustment in those at high risk of adverse events). Personalised medicine has made great strides especially in oncology, where an increasing number of therapies is used to target specific genetic alterations [1–4].

However, high prices are often charged for PM interventions [5, 6]. Manufacturers of PM have argued that their price setting is justified, as PM has benefits that are not captured in conventional health economic frameworks. Reimbursement authorities, however, have been hesitant to accept these claims. Additionally, they have in several cases rejected PM interventions for a lack of convincing evidence on treatment effectiveness [7]. Manufacturers have argued that the small patient populations that are inherent to PM (due to high levels of stratification) hamper the collection of data that meets current standards for health technology assessment (HTA) and have suggested that the solution lies

✉ Heleen Vellekoop
vellekoop@imta.eur.nl

1 Institute for Medical Technology Assessment, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands

2 Syreon Research Institute, Budapest, Hungary

3 Health Economics Research Centre, University of Oxford, Oxford, UK

4 Erasmus School of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

Key Points for Decision Makers

Certain features of personalised medicine (PM) have cast doubt on the appropriateness and applicability of the current health economic modelling toolkit. In line with previous research, our systematic literature review highlights considerable variation in methodological approaches to modelling PM.

To assist modellers of PM and evaluators of PM models, we developed comprehensive guidance with 23 recommendations on how to deal with the key modelling challenges in PM.

Eight recommendations are aimed toward the accurate modelling of testing pathways, given the importance of patient stratification in PM. Five recommendations deal with the estimation of treatment effectiveness, which may be complicated by the small patient groups in PM. Remaining recommendations discuss structural uncertainty and additional value elements, among other topics.

in updating HTA approaches. These (and other) issues have caused ambiguity about how to measure the value of PM, as reflected in the lack of national and international guidance on the evaluation of PM and considerable variation in the methodology and reporting in existing economic evaluations of PM [8, 9].

This study aimed to develop recommendations to health economic modellers in the field of PM and to those evaluating or reviewing PM models, to improve the consistency and quality across different health economic models of PM. The study was conducted within the context of the European Commission-funded Health Economics for Personalised Medicine (HEcoPerMed) Coordination and Support Action, which aims to identify optimal health economic modelling and payment strategies for evaluating and financing PM.

1.1 Working Definition of PM

The European Commission has defined PM as "a medical model using characterisation of individuals' phenotypes and genotypes (e.g. molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention" [10]. While this definition is comprehensive, it was deemed too broad for the purpose of this study, as many existing interventions—for which no new modelling challenges arise—could be argued to fall under this definition. Indeed, it has been argued that a new term such as PM should be used to describe new innovations that are

distinctive from well-established practices [11]. Additionally, while it is acknowledged that some interpret PM as improved stratification based on personal preferences and/or behaviour, it is mostly understood to be informed by biological information [11]. In line with this reasoning, the following definition of PM was adopted for this paper: "A medical model that bases therapeutic choice on the result of gene profiling or aims to correct pathogenic gene mutations".

Note that the decision *not* to treat is a therapeutic choice as much as the decision to treat is, and thus gene profiling that results in 'watchful monitoring' or 'no further medical treatment' is included in this definition. Furthermore, the profiling of gene mutations does not always require sequencing of the genes themselves. Profiling of gene mutations may be done at the functional level, for example, using protein expression tests.

2 Methods

Several research methods were used to develop the final recommendations.

2.1 Targeted Literature Review of Methodological Papers

First, a targeted review was conducted to identify methodological studies discussing challenges in the health economic modelling of PM. Methodological studies were identified through targeted searches on PubMed and Google Scholar (using search terms related to "methodology", "economic evaluation" and "personalised medicine"), through the scientific network of the authors, and by snowballing through the reference lists of identified studies. Only studies in English were considered for inclusion, with no limits on the publication year. The issues that were reported in the studies were extracted and combined into a list of methodological challenges. The list was used to develop questions for expert interviews.

2.2 Expert Interviews

The subsequent expert interviews aimed to gain a more current and in-depth understanding of the modelling challenges and to receive expert opinion on what constitutes good modelling practice. Eighteen experts were interviewed between November 2019 and February 2020, with all interviews lasting between 1 and 1.5 h. Interview candidates were identified through screening the authors of the studies identified in the targeted literature review. Interviewees were selected only if they had experience in developing, assessing and/or using economic models of PM. Experts were interviewed separately (though on one occasion two experts were

interviewed together) and interviews were conducted in a semi-structured manner. Interview responses were summarised per topic and reviewed for accuracy by the relevant interviewee. A data extraction table for the systematic literature review described below was developed based on the challenges identified in the targeted literature review and refined using the expert responses.

2.3 Systematic Literature Review of Economic Evaluations

A systematic literature review of economic evaluations of PM was conducted to identify methods that have been used to address the identified modelling challenges. The review was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. On 13 March, 2019, the following databases were searched: Embase, Google Scholar, MELDINE Ovid (similar to PubMed), and Web of Science. On 16 May, 2019, an additional search was performed in the Centre for Reviews and Dissemination and EconLit databases, as well as in the reimbursement dossier sections on the websites of the National Institute for Health and Care Excellence and the Institute for Clinical and Economic Review. The search strategy consisted of three groups of search terms, combined with the Boolean operator AND: "economic evaluation"; "modelling"; "personalised medicine". See the Electronic Supplementary Material (ESM) for the full search strategy. The search was limited to studies published in English from January 2009 onward. Given the rapid pace of innovation in the field of PM as well as improvements in HTA methodology over time, studies from before the 2009 cut-off were expected to be less relevant. Indeed, previous reviews have found increases in the quality and quantity of economic evaluations of PM since 2009 [4, 12]. Studies were only included if they presented a cost-effectiveness model incorporating final outcomes (i.e. life-years [LYs] or quality-adjusted life-years [QALYs]); met our working definition of PM; extrapolated outcomes beyond available clinical trial data; and described an existing (i.e. non-hypothetical) intervention. During the data extraction phase, the modelling methods used in the included studies were extracted if they addressed any of the challenges listed in the data extraction table.

2.4 Development of the Guidance

Finally, a detailed description of the challenges that may occur when modelling PM, as well as accompanying recommendations, were written based on the list of challenges identified in the targeted literature review, the summarised expert responses and the findings from the systematic literature review. The guidance was discussed at a stakeholder workshop and subsequently finalised. The stakeholder workshop took place in September 2020 and comprised around 30 participants, including health economists, representatives of national healthcare payers, and representatives from the pharmaceutical and diagnostics industry.

3 Results

3.1 Targeted Literature Review of Methodological Papers

Twenty-two methodological papers were identified through the targeted literature review (see the ESM for the full list). The papers discussed the challenges in the modelling and/or evaluation of interventions in a range of PM-related fields, such as genomic technologies [13], advanced therapeutic medicinal products [14] and gene therapies [15]. Several papers stated that the evaluation of PM is possible within existing HTA frameworks [15–18], though these are potentially not optimal [17]. One paper argued that changes to existing HTA methods and processes will likely be necessary as the field of PM develops [19], another proposed methodological adjustment [20]. The interview template that was developed based on the challenges reported in the methodological papers can be found in the ESM.

3.2 Expert Interviews

Several experts voiced similar opinions. Among these were the opinions that PM should be subject to the same methodological framework as other interventions to ensure consistency, and that additional value elements should generally not be included in economic evaluations until more research has been conducted and the consequences of including them are better understood. Many experts stated that a high degree of uncertainty is a key issue in PM. Uncertainty regarding treatment effectiveness (due to small study sample sizes) was considered especially prominent, as well as uncertainty whether the often-complex clinical pathways in PM are accurately reflected in economic models. Opinions differed regarding the desirability and necessity of using observational data to estimate treatment effectiveness. While some experts reasoned that economic evaluations based on observational data are unavoidable and/or acceptable, many experts argued that observational data are generally insufficient to inform pricing and reimbursement decisions and trial data should be demanded from manufacturers. Summarised responses, organised by topic, can be found in the ESM.

### 3.3 Systematic Literature Review of Economic Evaluations

The systematic literature review identified 7787 studies through database searching and seven additional studies by searching the reimbursement dossiers sections of the websites of the National Institute for Health and Care Excellence and Institute for Clinical and Economic Review. A total of 4774 individual studies remained after de-duplication [21], of which 195 were included after full-text screening (Fig. 1). The annual number of published economic evaluations of PM gradually increased over the search period, from ten studies in 2009 to 30 studies in 2018. More than half of the included studies (56%, $n = 109$) did not address any of the identified modelling challenges, while 33% ($n = 64$) of studies incorporated one or two aspects, and 11% ($n = 22$) incorporated three or more aspects. See the ESM for a list of identified studies, including an overview of the aspects incorporated in each study.

As shown in Table 1, the most addressed aspects were patients' treatment compliance (19%, $n = 38$) and uptake of testing (11%, $n = 22$), outcomes for relatives of index patients (14%, $n = 27$), and the conditionality of test sequences and results (9%, $n = 18$). The studies incorporating patient compliance and outcomes for relatives of index patients mostly concerned genetic testing for disease risk factors and subsequent preventive treatment. Uptake and compliance were generally incorporated as the proportions of patients undergoing testing and adhering to treatment. Outcomes for relatives of index patients were mostly combined with outcomes for index patients into single incremental cost and QALY figures, but sometimes reported separately. Studies incorporating conditionality of test sequences and results used a variety of assumptions, including varying sensitivity and specificity of tests across age groups, varying predictive values across ethnic groups, and conditional probabilities for test results (i.e. the probability of a positive/negative test result was dependent on whether a previous test had been positive/negative).

Studies incorporating additional value elements (6%, $n = 12$) mostly considered the psychological impact individuals experience when finding out about increased cancer risk through a genetic test (e.g. reduced uncertainty and/or increased anxiety), though two studies focused on psychological effects related to preventive surgery that may be performed based on the results of such a genetic test (e.g. reduced anxiety or worsened body image). These psychological effects were incorporated by applying a utility increase or decrease.

Fig. 1 Flow chart of the systematic literature review. *CE* cost-effectiveness, *ICER* Institute for Clinical and Economic Review, *NICE* National Institute for Health and Care Excellence, *PM* personalised medicine
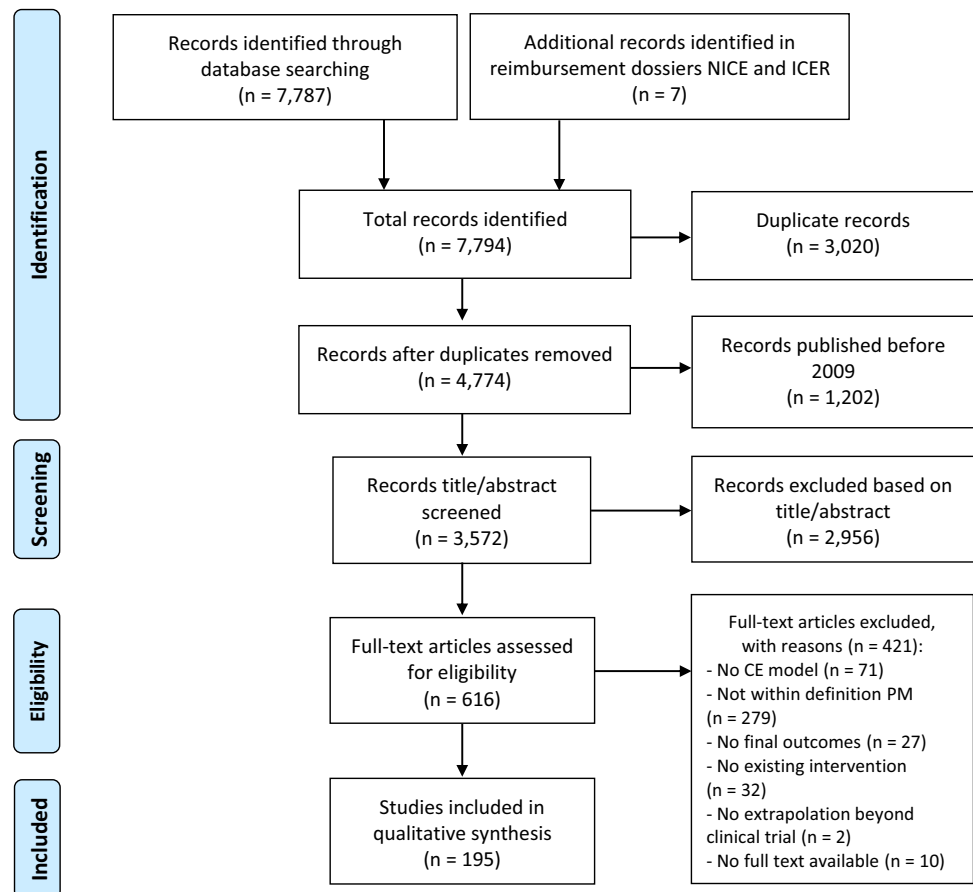
Studies that applied methods to account for potential bias in the extrapolation of outcomes for interventions with a proportion of long-term survivors (6%, $n = 11$) were mostly economic evaluations of chimeric antigen receptor (CAR) T-cell therapies. The methods included mixture cure models [22] and the use of hazard ratios to adjust general population mortality for long-term survivors of the condition in question.

Clinician compliance to protocols and guidelines ($n = 10$) and waiting times for patients ($n = 10$) were incorporated in 5% of studies. Relevant uncertainty analysis (4%, $n = 7$) evaluated, for example, the uncertainty around future cost reductions of genetic testing and the possible consequences of inconclusive genetic test results. The conditions of a managed entry agreement were incorporated in 4% ($n = 8$) of the studies, and methods to adjust for potential bias in non-RCT data in 3% ($n = 5$) of the studies. One identified method to adjust for potential bias in non-RCT data was to exclude patients with certain characteristics from the comparator cohort to increase comparability between the comparator and intervention cohorts. Another method consisted of the estimation of survival curves per treatment given and the subsequent weighting of the survival curves based on the expected distribution of treatments in the target population. In 2% ($n = 3$) of studies, lowered discount rates were applied for both costs and benefits. These studies evaluated gene therapies and cited the National Institute for Health and Care Excellence Methods Guide stating that a discount rate of 1.5% (instead of 3.5%) may be considered when a treatment restores individuals experiencing conditions with high

mortality and/or morbidity to sustained (near-)full health [23]. No studies parameterised expert judgement into a probability distribution.

3.4 Recommendations

The main outcome of this study is the following set of recommendations, informed by the results presented in the previous three sections. The numbered recommendations are introduced by an explanation. See the ESM to view the list of recommendations.

3.4.1 Perspective and Discounting

Some have argued that countries' standard HTA approaches may not always capture the full value of PM [20, 24, 25]. For example, for evaluating gene therapies that cure children from conditions with high mortality and morbidity, a societal perspective might be more appropriate than a payer perspective. That is, in addition to measuring patients' QALY gains, a societal perspective could capture the lifetime reduction in the use of informal care that might result from a child receiving a cure, as well as the potentially increased quality of life of carers.

Similarly, some have suggested a lower discount rate for health outcomes in PM with high upfront costs and long-term benefits, such as curative gene therapies and large-scale genetic screening programmes, to place a higher value on future benefits [14]. (A similar effect can be achieved with

Table 1 Results data extraction table

Topic	Aspect	Number (%) of studies addressing each aspect, $n = 195$
Discounting	Was there a deviation from standard discount rates for reasons particularly relevant to PM?	3 (2)
Test-treatment combinations	Was the conditionality of test sequences and results incorporated?	18 (9)
	Were waiting times incorporated?	10 (5)
	Were outcomes for relatives of index patients incorporated?	27 (14)
Effectiveness data	Were methods used to account for potential bias in non-RCT data?	5 (3)
Extrapolating outcomes for interventions with a portion of long-term survivors	Were methods used to account for potential bias in the extrapolation of outcomes for interventions with a portion of long-term survivors?	11 (6)
Additional elements of value	Were value elements beyond the QALY incorporated?	12 (6)
Incorporating compliance	Was patients' uptake of testing incorporated?	22 (11)
	Was patients' treatment compliance incorporated?	38 (19)
	Was clinicians' compliance to protocols and guidelines incorporated?	10 (5)
Uncertainty analysis	Were methods used to combine experts' judgements into a point estimate plus probability distribution?	0 (0)
	Was uncertainty analysis particularly relevant to PM performed?	7 (4)
Managed entry agreements	Were the conditions of a managed entry agreement incorporated?	8 (4)

PM personalised medicine, *QALY* quality-adjusted life-year, *RCT* randomised controlled trial

hyperbolic discounting, in which the discount rate is gradually reduced over time [26].)

However, allowing the assumptions regarding perspective and discounting to be different for (some) PM would hamper comparability of cost-effectiveness results across interventions and implicitly favour PM over other interventions, especially given the fact that many non-PM interventions also have wider societal and/or long-term benefits.

1. For economic evaluations of PM, use the standard perspective as recommended by national HTA guidelines in the base case.

2. For economic evaluations of PM, use the standard discount rates as recommended by national HTA guidelines in the base case.

3.4.2 Test-Treatment Pathways

Given that stratification is a key tenet in PM, testing plays an important role in the clinical pathway. A number of topics to consider when modelling tests are discussed below [27].

A single patient may be subject to a range of tests. There may be various options for combining the different tests. The tests may be performed in parallel or sequentially; when they are performed in sequence, choices may have to be made regarding the order in which the tests are performed. The diagnostic strategies chosen may vary across subgroups of the patient population. Additionally, some tests can be applied at different points in the treatment pathway (e.g. genome sequencing). As a result of these factors, modellers may be faced with many possible pathways. When prioritising which options to include in the model, a key consideration should be the extent to which they are relevant given the decision-making context.

3. Identify all relevant test-treatment pathways and justify why the pathways included in the model were selected.

When a test is used to stratify patients into subgroups that are eligible and non-eligible for a specific treatment, the consequences of using the test may affect the cost effectiveness of the test-treatment combination and should be explicitly considered. First, the costs of testing should be included in the economic evaluation of the treatment. When a new treatment requires the introduction of a new test (or the provision of an existing test to a wider target population), allocating 100% of the additional testing costs to the treatment under evaluation may be appropriate. This might seem unfair to "first movers", i.e. the first pharmaceutical or medical devices companies that require a specific test to be able to identify the right patients for their products, while the same test may later be used for other medical products. However, it is an accurate reflection of the decision problem at hand: the new treatment cannot be implemented in clinical practice without also implementing said test and thus their combined cost effectiveness should be assessed. When

the stratification for the new treatment can be done with a test that is already part of current practice, none or only a proportion of the testing costs may be allocated to the new treatment. The specific assumptions made regarding cost allocation may vary according to the budgeting and/or reimbursement arrangements in the decision-making context at hand. Note that total testing cost for all tested patients should be incorporated in the model, as opposed to the testing cost only for patients with a positive test result.

Furthermore, adverse events due to the testing procedure may reduce quality of life and increase mortality rates (e.g. a collapsed lung due to a lung biopsy). Additionally, the test results may stimulate further testing and treatment (e.g. because of secondary findings), affecting final health outcomes and costs. Additionally, there will be false-positive and false-negative patients among those tested, who may be facing poorer health outcomes, potentially leading to additional costs.

4. When a treatment requires the use of a test to stratify patients, include in the model the (downstream) costs and health outcomes of testing for both individuals who test (false-)positive and individuals who test (false-)negative.

The rates of false-positives and false-negatives are largely determined by the diagnostic accuracy of the testing technology used. The diagnostic accuracy of the technology is likely to vary according to the (subgroups of) the patient population in which the technology is applied and may change over time.

5. Ensure that the data used to estimate the diagnostic accuracy of a testing technology are appropriate to the patient population in the model.

Tests may have a continuous outcome, thus cut-off values must be set to determine the result (e.g. low, medium or high risk; positive or negative). Different cut-off values may be in use for the same test. For example, in the USA, pembrolizumab is indicated for patients with non-small cell lung cancer with high programmed death-ligand 1 (PD-L1) expression, which was first defined as PD-L1 expression in > 50% of tumour cells but later as PD-L1 expression in > 1% of tumour cells [28].

6. When different cut-off values are in use to determine test results, clearly define the cut-off value assumed in the base case. Investigate the effect of alternative cut-off values on cost-effectiveness results using a sensitivity analysis.

When various tests are modelled in sequence, their results may be correlated. See Box 1, for an example of the use of conditional probabilities to model test results.

7. When multiple tests are modelled in sequence, consider the interdependence between test results.

Patients presenting with symptoms generally do not receive treatment instantaneously. They may be faced with periods of waiting between presenting with symptoms and the decision to get tested, between the decision to get tested

The cost effectiveness of "expanded reflex" testing versus standard testing for HER2 mutations was estimated for breast cancer patients in the United States. Patients with HER2-positive breast cancer have a worse prognosis than HER2-negative patients but can be treated with trastuzumab, a monoclonal antibody specifically targeting HER2. HER2 status can be determined using either immunohistochemistry (IHC), which measures HER2 overexpression, or fluorescence in situ hybridisation (FISH), which measures HER2 gene amplification. In the standard approach, HER2 status is assessed using either IHC (80%) or FISH (20%). With expanded reflex testing, patients testing negative on either test subsequently receive the alternative test for confirmation. Interdependence between the two test results was incorporated in the model by using conditional probabilities obtained from a published systematic review. In the standard testing approach, only the outcome probabilities for either test in isolation are used. In the expanded reflex testing approach, the same outcome probabilities are used for the first test, while for the second test the outcome probabilities conditional on a negative outcome on the first test are assumed. Expanded reflex testing is estimated to render a QALY gain of 0.037 per patient treated compared to the standard approach, which is driven by a lower rate of (untreated) false negatives. Costs increase by $1455, resulting in an ICER of $39,721 per QALY gained. In the base case, it is estimated that 2.27% of patients are reclassified as HER2-positive after the second test. This percentage was varied between 1-8% in scenario analysis, rendering a maximum ICER of $47,110 and minimum ICER of $35,575.

Box 1 Considering interdependence between test results [29]

and testing, between testing and getting results, and between test results and the start of treatment. These waiting periods may impact outcomes, especially in conditions with high short-term morbidity/mortality (see Box 2 for an example).

8. If there is a notable risk of increased morbidity or mortality as a result of waiting periods, incorporate in the model the costs and health outcomes due to the waiting periods.

While some testing technologies are produced by a single provider and standardised, other tests are performed using local laboratory resources. There may be variation in testing costs between commercially developed test kits and local laboratory tests, as well as across laboratories.

9. Confirm that the assumed testing costs are accurate in the setting of interest and consider possible variation in costs across laboratories.

When an inheritable pathogenic mutation is identified in a patient, relatives are also at risk and may be offered genetic counselling and testing (e.g. in familial hypercholesterolaemia, BRCA-positive breast cancer). Focussing the economic evaluation of a test-treatment combination for inheritable mutations only on the index patients may offer an incomplete reflection of the clinical reality.

10. If relatives of index patients become eligible for genetic testing when the index patients test positive for a specific genetic marker, include the costs and health outcomes of testing relatives in the economic evaluation of the index patients.

3.4.3 Effectiveness Data

More stratification of patients leads to increasingly small patient (sub)groups, complicating the generation of (sufficiently) statistically powered data on treatment effectiveness through traditional randomised controlled trials (RCTs). Several alternative trial designs have been developed, including basket trials [31, 32], umbrella trials [31], n-of-1 trials [33] and adaptive trials [34]. While some of these alternative designs still allow for controlled studies, many are single-arm designs.

Doubts have been raised whether foregoing an RCT was justified in all of the non-RCT dossiers that have so far been submitted to regulators such as the European Medicines Agency and the US Food and Drug Administration [35]. Nonetheless, there appears to be increasing acceptance of non-RCT evidence among regulatory agencies [36].

The lack of RCT evidence poses challenges to health economic modelling as it complicates evidence synthesis (e.g. incomplete networks in a network meta-analysis) and increases the uncertainty around cost-effectiveness results. "Conditional reimbursement" or "coverage with evidence development" programmes, in which additional data are collected after market approval, have raised concerns. First, it has been questioned whether they are able to provide unbiased estimates of relative effectiveness when relying on observational data [37]. Concerns also exist about the feasibility of withdrawing medicines that were granted

The cost effectiveness of tisagenlecleucel, a CAR T-cell therapy, was estimated for paediatric patients with relapsed/refractory B-cell acute lymphoblastic leukaemia in the United States. Patients undergo leukapheresis so that their T-cells can be harvested and used to prepare the tisagenlecleucel infusion. During the waiting period between leukapheresis and receiving the infusion, patients are at risk of mortality. This is captured in the model by a short-term decision tree, in which the first node reflects the probability that patients receive the tisagenlecleucel infusion. The chance node has the branches "continue with the infusion" and "die before receiving the infusion". An additional branch captures the probability that patients "discontinue before infusion because of adverse events or manufacturing failure". In a scenario analysis where the model did not incorporate waiting time by starting at infusion instead of at leukapheresis, the incremental costs increased from $329,498 to $454,892 and incremental QALYs increased from 7.2 years to 9.1 years. The increase in incremental costs and QALYs is explained by the exclusion of patients who do not receive the tisagenlecleucel infusion and therefore have both lower treatment cost and shorter quality-adjusted life expectancy. Excluding waiting time from the model slightly increased the incremental cost-effectiveness ratio, from $46,000 per QALY to $50,000 per QALY.

Box 2 Incorporating waiting times [30]

reimbursement once further evidence does not demonstrate their (cost-)effectiveness [38, 39].

11. Where possible, use effectiveness data from trials with two (or more) alternative treatment strategies.

In tandem with the increasing acceptance of non-RCT evidence, there has been a rise in the use of evidence from early trials [40]. However, the relationship between final outcomes and the surrogate outcome measures often used in early trials is not always well established. For example, out of 93 cancer drug indications for which accelerated approval was granted by the US Food and Drug Administration based on surrogate outcomes (such as response rate or progression-free survival), confirmatory trials reported improved overall survival for only 19 (20%) [41].

12. When surrogate outcomes are used to estimate final outcomes, specify which data sources were used to estimate the relationship between surrogate and final outcomes and justify any assumptions made about the relationship.

When only data from single-arm studies are available, external data could be used to construct a control arm. However, treatment effectiveness may improve over time, owing to treatment-related factors such as dosing optimisation, improvements in the standard of care or external factors such as improvements in general population health. Historical data may, therefore, underestimate effectiveness in the control group, leading to the overestimation of the new treatment's effectiveness.

13. When effectiveness of the comparator is estimated using external data, account for a possible time trend in effectiveness.

Increasingly many treatments that target specific genetic markers are coming onto the market. If the genetic marker affects disease prognosis, combining data sources with a different prevalence of the genetic marker may be inappropriate for estimating comparative effectiveness. Nonetheless, the effectiveness estimate can potentially be adjusted when the prognostic value of the marker and the prevalence of the genetic marker across the different data sources are known.

For example, two TRK inhibitors (larotrectinib and entrectinib) have recently come onto the market for tumours with NTRK gene fusions. The effectiveness of both treatments was assessed using single-arm trials, meaning that external data are necessary to be able to construct a comparator arm reflecting the standard of care. However, while the trial population for the TRK inhibitors consisted of only NTRK-positive patients, the available data on comparator effectiveness stem from trials in which most patients were NTRK negative, owing to the low prevalence of NTRK fusions in many types of cancer. Preliminary evidence suggests that NTRK-positive patients have a worse prognosis [42]. The treatment effectiveness estimated on populations with a large share of NTRK-negative patients may therefore provide a biased estimate of the treatment effectiveness for NTRK-positive patients and may have to be adjusted.

14. When effectiveness of the comparator for patients with a specific genetic marker is estimated using external data, account for the prognostic value of the genetic marker and differences in its prevalence across the different data sources.

When new gene tests are developed to allocate patients to existing treatments, it is unlikely new RCTs will be

performed for the existing treatments in each of the subgroups introduced by the new test. Instead, data on treatment effectiveness in the subgroups may come from genotype-phenotype studies, in which associations between genetic markers and clinical outcomes are investigated. Estimating a causal relationship between genotype and phenotype tends to be highly complicated, owing to, among other things, the potential for gene–gene and gene–environment interactions, heterogeneity in genetic markers (e.g. hundreds of different BRCA mutations have been found in patients with breast cancer [43]) and heterogeneity in clinical symptoms [44]. Additionally, it may be difficult to identify appropriate controls, to link genetic data to data on clinical outcomes [45] and to get a large enough sample size to meet statistical significance requirements. As a result, details about the relationship between a genetic marker and clinical outcomes are often uncertain or unknown. For example, while it has been shown that patients with acute coronary syndrome who carry a loss-of-function polymorphism of cytochrome P450 2C19 (CYP2C19) experience more thrombotic events when treated with clopidogrel, uncertainty remains regarding the degree of association between carrier status and thrombotic events [46]. Similarly, the relationship between the level of HER2 expression in patients with breast cancer and the extent to which progression-free survival is reduced is not fully known [47].

15. Specify which data sources were used to estimate the association between the genetic marker(s) of interest and clinical outcomes and justify any assumptions made about the association.

3.4.4 Extrapolating Survival

So far, innovations in PM have mostly been in disease areas with high mortality, such as oncology and rare severe genetic disorders. An accurate estimation of the effect that these interventions have on patient mortality is key to assessing their cost effectiveness. Trials generally provide only short-term data, bringing the need for modelling to estimate survival beyond the trial period. While long-term survival can sometimes be estimated using the surrogate outcomes measured in the trial (see recommendation 12), it is more common to extrapolate from the short-term mortality captured in the trial. The choice of survival model is often informed by assessing the statistical fit to the data. However, models with a good fit to short-term trial data do not always provide plausible predictions regarding long-term outcomes [48]. Expert judgement may be used to evaluate the plausibility of the estimated survival.

16. When extrapolating survival data beyond the study period, use expert opinion alongside statistical fit to choose the survival model.

In several PM interventions, patients might be considered "cured" when they experience long-term survival. This may apply to, for instance, oncology patients experiencing sustained complete remission after receiving targeted therapy, or to patients with early-onset diseases with high mortality who respond to a gene therapy. However, even if patients experience long-term survival, they may be faced with poorer long-term health outcomes than the general population. For example, Janssen-Heijnen et al. showed that for several cancer types (stomach cancer, non-small cell lung cancer, stage II or III breast cancer, prostate cancer, Hodgkin lymphoma), patients who had been "cured" and survived between 10 and 20 years after diagnosis still had poorer survival than the general population [49]. They hypothesised that this could be due to late recurrences, secondary tumours or comorbidities associated with cancer risk factors [49]. Assuming (age- and sex-specific) general population mortality for "cured" patients may therefore be inappropriate (see Box 3 for an example).

17. When extrapolating survival data beyond the study period, account for any excess mortality and morbidity among long-term survivors.

3.4.5 Additional Elements of Value

It has been argued that the QALY insufficiently captures the full value interventions may have. The ISPOR Value Assessment Framework Special Task Force identified a list of additional value elements to be included in a cost-effectiveness analysis, including scientific spill-overs, equity, real option value, value of hope, severity of disease, insurance value, fear of contagion and reduction in uncertainty [20]. A related concept that has been suggested is "personal utility", which is generally used either to describe the value of knowledge (e.g. knowledge of a test outcome) or as an umbrella term for the non-health outcomes that individuals might value [51, 52]. Patients may indeed value outcomes of healthcare beyond increased health. Diagnostic information, for example, may allow patients to make better life decisions or cause psychological effects such as alleviated (or increased) anxiety [53]. However, the suggested additional value elements raise several concerns.

First, it remains unclear how to define, measure and value many of the identified elements, partly owing to their conceptual ambiguity. There appears to be a risk of double counting, both within the set of elements (e.g. *severity of disease* can be argued to be part of *equity*; *insurance value* is likely strongly correlated with *severity of disease*, given that the value of being insured against the consequences of falling ill is higher when the diseases covered are more severe) and between the elements and the QALY (some of the "additional" elements may already be captured in

In a report by the Institute for Clinical and Economic Review on the cost effectiveness of CAR T-cell therapies in B-cell malignancies, tisagenlecleucel and axicabtagene ciloleucel were evaluated, respectively for paediatric patients with relapsed/refractory B-cell acute lymphoblastic leukaemia and adults with relapsed/refractory aggressive B-cell lymphoma. Survival was extrapolated from the available trial data, owing to the limited observation period in the clinical trials (15.4 months in the ZUMA-1 trial, 3.7 months in the JULIET trial). It was presumed that patients who survived to the point at which the survival curve flattened (i.e. the slope equalled zero) were among the long-term survivors. A knot was introduced at this point in the survival curve (at five years), after which patients were assumed to be subject to general population mortality, adjusted for excess mortality observed among long-term survivors of B-cell malignancies. Based on published evidence, a standardised mortality ratio of 9.1 was used to reflect excess mortality among long-term survivors in the paediatric leukaemia cohort, while a standardised mortality ratio of 1 was used for the adult B-cell lymphoma cohort. The latter proved influential on model outcomes in one-way sensitivity analysis. The base case assumption of a standardised mortality ratio of 1 rendered an incremental cost-effectiveness ratio of $136,078 for axicabtagene ciloleucel versus chemotherapy, which increased to $176,491 when a maximum standardised mortality ratio of 3.4 was assumed. An additional scenario analysis was conducted, in which it was assumed that only 80% of those alive and responding to treatment at five years would be long-term survivors in the paediatric leukaemia cohort and 95% of those alive and responding to treatment at five years would be long-term survivors in the adults B-cell lymphoma cohort. The remaining patients (20% and 5%, respectively) were assumed to die at five years. This increased the incremental cost-effectiveness ratio for tisagenlecleucel versus clofarabine from $45,871 to $53,195 and the incremental cost-effectiveness ratio for axicabtagene ciloleucel versus chemotherapy from $136,078 to $140,443.

Box 3 Accounting for excess mortality among long-term survivors [50]

preference-based quality-of-life assessments, such as the *reduction in uncertainty*).

Additionally, there appears to be a focus on positive value elements, while negative value elements may be equally relevant. For instance, the value of hope that patients might experience prior to treatment (e.g. the hope that they are among the long-term responders to treatment) may be (partly) offset by the disutility due to dashed hope once treatment outcomes are known.

It is important to be aware that including additional value elements may alter decision making, at the expense of length and quality of life. For example, once additional value elements are included, intervention A with high health benefits might be deemed less cost effective than intervention B with medium health benefits but many additional elements of value. When choosing to adopt intervention B instead of intervention A, we are implicitly trading off length and quality of life against other value elements (see Table 2 for a stylised example).

Individuals may indeed be willing to trade off certain elements of value against length and quality of their own life [54], conveying consumer value. However, this does not

necessarily mean that people are willing to make such trade-offs between, for example, hope and health in others, casting doubt on whether such value elements should be prioritised at the national level. Indeed, it may be debated whether healthcare payers such as national governments should pay for all elements that bring value to individual patients, especially given that there likely is significant variation across patients in their valuations of specific elements, as well as within patients depending on the time they are asked. For example, risk-averse patients might not experience any value of hope and risk-loving patients may only experience a value of hope for a short time, after which dashed hope might be experienced.

If additional value elements were to be included in economic evaluations, this should be done for all interventions, not just PM, to ensure consistency and comparability across studies. Indeed, the suggested value elements may be relevant outside of PM. For example, while a patient with a family history of breast cancer likely experiences relief (i.e. gains personal utility) when they find out they are BRCA negative (PM), personal utility may equally be gained by a

Table 2 Stylised example of the consequence of including a value of hope

Treatment	Cost-effectiveness threshold (λ)	ΔQALY	ΔCost	Value of hope	Incremental net monetary benefit[a]
Treatment 1 (standard approach)	$50,000	2	$80,000	–	$20,000
Treatment 2	$50,000	2.5	$80,000	0	$45,000
Treatment 1 (including value of hope)	$50,000	2	$80,000	$30,000	$50,000[b]

QALY quality-adjusted life-year

[a]Incremental net monetary benefit $= (\lambda \times \Delta$QALY$) - \Delta$cost

[b]Incremental QALYs are higher for treatment 2. However, treatment 1 offers "hope", while treatment 2 does not. In this example, placing a monetary value on "hope" increases the incremental net monetary benefit of treatment 1 from $20,000 to $50,000. This may lead to the prioritisation of treatment 1 over treatment 2, despite treatment 2 offering higher QALY gains

couple wanting to get pregnant when they find out they have no fertility issues (non-PM).

Note that the threshold value against which the cost effectiveness of the intervention is judged may have to be adjusted to account for the additional value elements. For example, if additional value elements are included in a sensitivity analysis, the resulting cost-effectiveness outcomes may have to be judged against a different threshold than the outcomes in the base-case analysis. The rationale for this depends on whether the threshold is viewed as a supply-side estimate (i.e. the opportunity cost of healthcare spending, or the marginal productivity of the healthcare system) or a demand-side estimate (i.e. the societal willingness to pay for improvements in health) [55]. In the former, the threshold changes when additional value elements are included because the opportunity cost of spending now includes not only health forgone but other benefits forgone as well. In the latter, the threshold changes because the societal willingness to pay for only health outcomes may be different from the willingness to pay for health and non-health outcomes combined.

18. Only include elements of value recommended by national HTA guidelines in the base case. If additional elements of value are included in a sensitivity analysis, ensure possible elements of negative value are equally considered and included for both the intervention and the comparator.

3.4.6 Incorporating Compliance

While health economic models are used to simulate the clinical reality, clinical reality is not always optimal. Depending on the decision context, modellers may choose to model a healthcare intervention at its optimal implementation level or at a level of implementation that is closer to reality (or both). It is important to be transparent about the extent to which the model reflects optimal implementation.

In PM, a significant cause of suboptimal implementation may be imperfect compliance owing to the perceptions and preferences of patients and clinicians regarding PM. For example, unwillingness to find out about risk-increasing gene mutations may hamper patients' uptake of genetic testing. Similarly, limited understanding of risk/probabilities may lessen patients' compliance to therapeutic plans based on genetic testing. In addition, clinicians may not be fully compliant to protocols and guidelines because of a limited knowledge of genetics, or they may already initiate treatment in rapidly deteriorating patients if the waiting time for test results is too long.

Compliance is likely affected by the perceived probability of disease (this applies to testing only), the severity of disease and/or the type of treatment (e.g. preventive or curative). For example, in the study described in Box 4, the uptake of genetic testing for the risk of breast and ovarian cancer is markedly lower for people aged under 30 years than for people aged 30 years and above. Similarly, compliance to genetic testing for cardiovascular disease risk and subsequent preventive measures may be lower than compliance to genetic testing of tumours and subsequent cancer treatment.

When incorporating compliance, note that an adjustment of the effectiveness estimate might not be necessary for pragmatic trials, where data reflect real-world compliance. Furthermore, note that reduced compliance does not automatically mean that intervention costs are lower (e.g. medicines may have been dispensed but not taken). Finally, patient compliance may vary considerably between for example socioeconomic, geographic and age groups. Clinician compliance might also vary according to the societal group their patients belong to.

19. Include parameters reflecting patient and clinician compliance in economic evaluations for decision makers who require cost-effectiveness results under realistic circumstances.

20. When including patient and clinician compliance in economic evaluations, confirm that the assumed compliance is accurate in the setting of interest and consider possible variation in compliance across societal groups.

The cost effectiveness of genetic testing was evaluated for patients with early-onset breast cancer and their female relatives in Norway. BRCA-testing is currently routine care, with identified carriers being at higher risk of breast and ovarian cancer and being offered risk-reducing surgery (bilateral mastectomy with reconstructive surgery and bilateral salpingo-oophorectomy). Alternative 7- and 14-gene panels, which include BRCA1/2 as well as additional genes associated with increased risk of breast and ovarian cancer, were compared against BRCA testing only. Both *imperfect uptake of testing* and *imperfect uptake of risk-reducing surgery* were incorporated in the model. The input values for the uptake parameters were obtained from observational data and assumed to have a beta distribution. Uptake of testing as well as risk-reducing surgery among relatives of index patients were assumed to vary with age. Mean uptake of genetic testing among relatives of carriers was assumed to be 0.30 for ages 18-29; 0.82 for ages 30-49; and 0.80 for ages ≥ 50. Mean uptake of prophylactic mastectomy was assumed to be 0.12 for ages 25-34 and 0.11 for ages 35-60. Mean uptake of prophylactic salpingo-oophorectomy was assumed to be 0.10 for ages 25-34; 0.28 for ages 35-39; and 0.35 for ages 40-60. Uptake of the risk-reducing surgeries was assumed to be higher among index patients than their relatives (mean uptake of 0.39 for mastectomy among index patients, 0.36 for salpingo-oophorectomy). The incremental cost-effectiveness ratio of the 7-gene panel versus BRCA only was estimated to be $53,310, while the incremental cost-effectiveness ratio of the 14-gene panel versus the 7-gene panel was estimated to be $127,071.

Box 4 Incorporating compliance [56]

3.4.7 Uncertainty Analysis

To enable optimal reimbursement decisions, it is important to present uncertainty in the cost-effectiveness estimates. Personalised medicine tends to be rife with parameter and structural uncertainty. While PM is often associated with the tailoring of treatments to individual patients, in practice, PM generally divides patients into groups, albeit small groups, and provides the same treatment within each group. The small sample sizes in RCTs and the use of observational data tend to increase the uncertainty of the treatment effect. Other input parameters in economic models of PM may also be uncertain, such as the prevalence of the genetic marker in the target population, testing costs, and the sensitivity and specificity of the testing technology used.

Given limited data availability, expert judgement may be used to provide estimated values for the input parameters. However, expert judgement, too, carries uncertainty, which should be reflected. Several methods have been developed to synthesise the estimated values by multiple experts for a single parameter into a probability distribution that can be included in a sensitivity analysis [57–61].

21. When expert judgement is used to estimate values for the input parameters in the model, synthesise the elicited values into a probability distribution to be included in a sensitivity analysis.

Considerable structural uncertainty arises in PM owing to, for example, the myriad assumptions and decisions that must be made about how to reflect complex testing and treatment pathways, about the expected duration of treatment effect or regarding the methods used to obtain effectiveness estimates when RCT data are not available. Structural uncertainty may have a significant impact on cost-effectiveness results. A failure to assess structural uncertainty provides an incomplete depiction of the decision problem to decision makers. Structural uncertainty may be assessed through a sensitivity analysis in which the effect of plausible alternative assumptions and decisions is investigated. However, performing a sensitivity analysis does not allow for the assessment of the decision uncertainty and the value of information [62]. Alternative options are (i) model averaging, which can provide an assessment of decision uncertainty and (ii) the parameterisation of structural uncertainties, which allows for the assessment of both decision uncertainty and the value of information [62].

22. Identify uncertainties in structural assumptions and decisions and investigate their impact on cost-effectiveness results through a sensitivity analysis. Parameterise structural aspects where possible.

3.4.8 Managed Entry Agreements

When uncertainty regarding an intervention's comparative effectiveness, cost effectiveness and/or budget impact precludes a reimbursement decision, as may commonly be the case for PM, managed entry agreements (MEAs) between

The cost effectiveness of tisagenlecleucel was estimated for paediatric patients with relapsed/refractory B-cell acute lymphoblastic leukaemia in the United States. At the time of the study, the manufacturer of tisagenlecleucel was using an outcome-based payment scheme whereby the payer only bears the cost of treatment when the patient achieves initial remission (i.e. no cure, no pay). It was found by the authors that the cost effectiveness of tisagenlecleucel under the existing outcomes-based payment scheme was similar to the cost effectiveness without such a payment scheme, owing to the high rate of initial remission among patients treated with tisagenlecleucel. In a scenario analysis, the effect of changing the conditions of the payment scheme was evaluated. Under a payment scheme whereby the payer bears the cost of treatment only if the patient is still in remission at 6 months, tisagenlecleucel's incremental cost-effectiveness ratio improved from $74,000 (under the assumption of 40% relapse-free survival; $188,000 under the assumption of 20% relapse-free survival) to $47,000 ($115,000). When the payer was assumed to bear the cost only if the patient is still in remission at 12 months, the incremental cost-effectiveness ratio improved to $28,000 ($58,000).

Box 5 Incorporating the conditions of a managed entry agreement [64]

a healthcare payer and a manufacturer can be used to offer patients access to the intervention. The two main types of MEA are financial (e.g. discounts, price-volume agreements) and outcomes based (e.g. payment based on individual patient response) and both can be constructed with different levels of risk sharing [63]. The conditions of a MEA can be incorporated into health economic models, as some of the costs may be shifted to a different point in time and should be discounted appropriately. Indeed, models may be used to optimise the conditions of a MEA, such as the period a pharmaceutical and/or diagnostic test is provided at no cost, the price cap or the time point at which treatment effectiveness for individual patients is assessed (see Box 5 for an example). The optimisation criterion could be a combination of discounted cash flow (most relevant from the manufacturer's perspective) and incremental net monetary benefit and budget constraints (most relevant for payers).

23. If a managed entry agreement is being considered for an intervention, include its conditions in the model evaluating the intervention.

4 Discussion

Twenty-three recommendations for economic evaluations of PM are provided in this study, covering a broad range of topics. As mentioned in Sect. 1, the recommendations were developed against a backdrop of calls for the review of HTA methodology, given developments in the field of PM. A systematic and multi-faceted effort was therefore undertaken to assess the extent to which existing HTA methods need to be adapted for PM. The consensus among interviewed experts was that existing methods are adequate and appropriate for assessing PM and non-PM alike. The experts also felt that, within jurisdictions, PM interventions should be subject to the same basic HTA framework as non-PM interventions, to ensure comparability between economic evaluations and consistency in decision making (as reflected in recommendations 1 and 2). Nonetheless, several challenges were identified that may be faced by those producing or evaluating economic evaluations of PM.

The guidance aims to serve as an overview of topics that should be considered for economic evaluations of PM. Some of the recommendations may remind modellers and evaluators of good practices that are often neglected (e.g. recommendations 3–5), others may provide direction when modellers and evaluators are uncertain how to proceed in the face of ongoing debate (e.g. recommendation 18). The guidance is intended to be used in addition to, rather than as a replacement of, existing, more general modelling guidance [65–69].

It is acknowledged that the recommendations are not relevant exclusively to PM as several challenges in the economic evaluations of PM are also encountered in non-PM. For example, the issue of large upfront costs with benefits stretching far into the future (mentioned under Sect. 3.4) also appears in the modelling of vaccination programmes. Nonetheless, PM is unique in the range and extent of challenges it faces. Our recommendations are therefore particularly valuable in the modelling of PM.

Last, it should be noted that it is unlikely that all recommendations are relevant to each economic evaluation of PM, and some may be not feasible because of limited data and/or research resources. It is therefore left to the modeller to disregard recommendations when appropriate, though a justification for doing so should be provided.

4.1 Limitations

As mentioned in Sect. 1, there is a range of interpretations of the term "personalised medicine" and the working definition of PM in this paper may not capture all of them. As a result, the developed guidance might not completely meet the needs of those with a different understanding of PM. Nonetheless, the definition of PM adhered to in this study focuses on "new innovations", which are those most likely to require additional modelling guidance. Those who hold a definition of PM that is more inclusive of well-established healthcare may find that existing guidance suffices for these interventions. Further work may be needed to meet the needs of those who understand PM to be informed by patient preferences [70].

Although machine learning-based technologies and digital health applications are sometimes classed under PM (e.g. [19]), they were excluded from our working definition of PM, for two main reasons. First, it may be debated to what extent these technologies are "personalised". Many machine learning approaches are rooted in statistics. There are numerous statistical tools that are widely used in medicine (such as the Simple Calculated Osteoporosis Risk Estimation prediction model) and not necessarily seen as "new innovations" or as PM. This begs the question as to where to draw the line between various statistical models in deciding whether they can be classed under PM. Digital health applications are often used to complement or automate existing in-person healthcare services. They may for example be used to send automated appointment reminders to patients, to enable online consults with physicians, to automate some of the administrative tasks for healthcare professionals, to capture patients' health data or to monitor patients from a distance. While these developments mark a shift in the mode of healthcare delivery, they are not clearly more "personalised" than the existing healthcare practices that are usually not regarded as PM. Second, although search terms related to "machine learning" and "digital health" were included in the literature searches, it rendered little relevant hits. Most studies in the digital health category considered relatively simple devices for the monitoring of blood glucose in patients with diabetes mellitus, while no studies were identified for the machine learning category. It was therefore decided that insufficient literature on the economic evaluation of machine learning-based technologies and digital health applications was available to allow for their inclusion in this study.

A certain degree of interpretation and subjective prioritisation of the research findings was inevitable in developing the recommendations, given the normative nature of guidance on what constitutes "good practice". This issue was inherent to the research goal and was mitigated by the fact that voices of many different perspectives, backgrounds, countries and types of expertise were heard, both during the expert interviews (18 experts from different backgrounds and with different specialisations were interviewed), the drafting of the guidance (the recommendations were based on consensus opinion within the sizeable HEcoPerMed consortium) and the stakeholder workshop (around 30 participants from various fields were present).

4.2 Implications

A substantial amount of literature on the health economic modelling of PM already exists. Several studies discuss challenges in the modelling of PM and suggest potential solutions but do not provide clear guidance to health economic modellers and/or the evaluators of health economic models [14, 71–73]. Other relevant studies do provide guidance, mostly in the form of a quality checklist, but on topics more specific than "personalised medicine". Among these are checklists by Kip et al. [27] and Yang et al. [74] to assess the quality of economic evaluations of diagnostic tests, a checklist to assess economic evaluations of gene therapies [15], and a checklist of PM models focusing on the need for patient-level modelling [75]. A study by Christensen et al. [76] provides some guidance specifically on how to measure the costs of integrating whole-genome sequencing (WGS) into clinical practice.

Note that no recommendation on the specific modelling technique to be used in PM was included in this study as the appropriate modelling technique was deemed to depend on the decision context. Nonetheless, it is acknowledged that the rise of PM has been argued to call for a more widespread use of patient-level modelling (as opposed to cohort-level modelling, which is currently most prevalent), owing to the former's ability to simulate a greater variety of clinical pathways and easily include patient history into the analysis [75, 77]. In patient-level modelling, in addition to considering the parameter and structural uncertainty that are discussed in the guidance, special attention should be paid to addressing patient heterogeneity and stochastic uncertainty [78].

The results of this systematic review add to the existing literature by providing comprehensive and specific guidance to all modellers of PM and evaluators/reviewers of PM models. This may increase the consistency, comparability and quality of economic evaluations of PM, and therefore improve the evidence about the added value of PM. The guidance could also be used for developing and/or evaluating early health economic models, though several recommendations may be more difficult to implement because of limited data (e.g. recommendation 11).

Even for standard health economic models, various recommendations encourage the use of data that may not be available in practice, among which: effectiveness data obtained through trials with two (or more) treatment strategies (recommendation 11); data on the relationship between surrogate outcomes and final outcomes (recommendation

12); and data on the prognostic value of the genetic marker of interest (recommendation 14). The absence of these data items can hamper an accurate assessment of the cost effectiveness of PM, as it may introduce a high level of uncertainty regarding the cost-effectiveness results. Although marketing authorisation may be granted to pharmaceuticals based on relatively limited data, the above-mentioned recommendations highlight that data needs are different for health economic modelling. This is in line with recent examples of PM receiving approval from the European Medicines Agency but subsequently being rejected for reimbursement by national HTA bodies because of inconclusive evidence [7]. While modelling can be used to address some of the data limitations, the main solution to insufficient data may be increased communication between regulators and national HTA agencies about what type of data is needed.

Recommendation 4 urges modellers to include downstream costs and health outcomes of testing. While estimating downstream costs and health outcomes is likely to be feasible for targeted gene panels, which are currently most widely used in healthcare, it may become an increasingly unwieldy task as whole-exome sequencing and WGS become more ingrained in standard clinical practice. Whole-exome sequencing and WGS can find genetic variants associated with (increased risk of) a wide range of conditions. Estimating the effect of the identification of these variants and subsequent (preventive) actions therefore requires knowledge of and data from many disease areas. Existing studies on the topic have tended to simplify their analyses, for example, by only considering short-term downstream consequences [76] or a subset of possible disease areas [79]. More research on how to best capture downstream costs and health outcomes may be valuable. Nonetheless, estimates of the downstream costs and health outcomes of whole-exome sequencing and WGS are bound to be shrouded in uncertainty for the foreseeable future as much is still unknown about the relationships between genotype and phenotype. That is, whole-exome sequencing and WGS may identify many variants of *unknown* significance to the person's health, severely hindering a reliable estimation of the downstream costs and health outcomes of applying these technologies. The solution to this may lie to a larger extent in continued biomedical research rather than in increasingly sophisticated HTA methods.

In Sect. 3.4.5, it was argued that it is unclear how to measure many of the suggested value elements, partly owing to their conceptual ambiguity. It was also noted that there appears to be an unduly focus on positive value, with limited attention for plausible elements of negative value. Recommendation 18 therefore discourages the incorporation of additional value elements in base-case analyses. Nonetheless, it is acknowledged that value elements beyond the traditional QALY may be relevant in decision making. Indeed, some elements, such as equity, are routinely considered in some countries. Currently, additional value elements tend to be incorporated qualitatively (though numerical values are sometimes used in a multiple-criteria decision analysis [80]), which does not enable the explicit assessment of the trade-offs between length and quality of life on the one hand and other value elements on the other hand. Quantifying additional value elements (in combination with estimating the change in the cost-effectiveness threshold if additional value elements are included) would provide insight into these trade-offs. Further research may be conducted to identify all relevant elements of value, clearly define them and determine how to measure them. Nonetheless, researchers are encouraged to stay mindful of the difference between value to individuals and value to society at large, given that healthcare payers often make decisions at the societal level and may not be willing to pay for all elements that bring value to individual patients.

5 Conclusions

This study provides a comprehensive list of recommendations to modellers of PM and evaluators/reviewers of such models. The recommendations provide valuable guidance, given the ongoing discussions about the value of PM and the many modelling complexities brought about by PM, and aim to contribute to improved consistency and quality across different health economic models of PM.

Declarations

Conflicts of interest/competing interests Heleen Vellekoop, Simone Huygens, Matthijs Versteegh, László Szilberhorn, Tamás Zelei, Balázs Nagy, Rositsa Koleva-Kolarova, Apostolos Tsiachristas, Sarah Wordsworth and Maureen Rutten-van Mölken have no conflicts of interest that are directly relevant to the content of this article.

Ethics approval According to the Dutch Law on Medical Scientific Research (WMO), this study did not require approval by a medical ethics committee.

Consent to Participate All interviewed experts consented to participation in the study.

Consent for Publication All interviewed experts consented to their statements being incorporated in the study's findings in an anonymised manner.

Code Availability No programming was used in this study.

Availability of data and material Various data items are made available through the supplementary files, including the search strategies used for the systematic literature review; the studies identified through both of our literature searches; the interview template for the expert interviews and summarised responses from the expert interviews. The Excel file used for the data extraction during the systematic literature review can be made available upon request.

Authors' contributions MV and MR outlined the concept and broad methods for this study. HV, SH, MV and MR conducted the two literature reviews and the expert interviews. HV, SH, MV and MR developed the first set of recommendations, which was critically revised by all authors. LS, TZ and BN organised the stakeholder workshop. HV wrote the manuscript. All authors provided multiple rounds of critical feedback.

References

1. D'Andrea E, Marzuillo C, Pelone F, et al. Genetic testing and economic evaluations: a systematic review of the literature. Epidemiol Prev. 2015;39:45–50.
2. Wright SJ, Newman WG, Payne K. Accounting for capacity constraints in economic evaluations of precision medicine: a systematic review. Pharmacoeconomics. 2019;37:1011–27.
3. Hatz MH, Schremser K, Rogowski WH. Is individualized medicine more cost-effective? A systematic review. Pharmacoeconomics. 2014;32:443–55.
4. Plothner M, Ribbentrop D, Hartman JP, et al. Cost-effectiveness of pharmacogenomic and pharmacogenetic test-guided personalized therapies: a systematic review of the approved active substances for personalized medicine in Germany. Adv Ther. 2016;33:1461–80.
5. Ferkol T, Quinton P. Precision medicine: at what price? Am J Respir Crit Care Med. 2015;192:658–9.
6. O'Sullivan BP, Orenstein DM, Milla CE. Pricing for orphan drugs: will the market bear what society cannot? JAMA. 2013;310:1343–4.
7. Touchot N, Flume M. Early insights from commercialization of gene therapies in Europe. Genes (Basel). 2017;8:78.
8. Vegter S, Boersma C, Rozenbaum M, et al. Pharmacoeconomic evaluations of pharmacogenetic and genomic screening programmes: a systematic review on content and adherence to guidelines. Pharmacoeconomics. 2008;26:569–87.
9. Beaulieu M, de Denus S, Lachaine J. Systematic review of pharmacoeconomic studies of pharmacogenomic tests. Pharmacogenomics. 2010;11:1573–90.
10. European Commission. Commission Staff Working Document. Brussels, Belgium. https://ec.europa.eu/research/health/pdf/2013-10_personalised_medicine_en.pdf. Accessed 27 June 2019
11. Schleidgen S, Klingler C, Bertram T, et al. What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC Med Ethics. 2013;14:55.
12. Simeonidis S, Koutsilieri S, Vozikis A, et al. Application of economic evaluation to assess feasibility for reimbursement of genomic testing as part of personalized medicine interventions. Front Pharmacol. 2019;10:830.
13. Buchanan J, Wordsworth S, Schuh A. Issues surrounding the health economic evaluation of genomic technologies. Pharmacogenomics. 2013;14:1833–47.
14. Jonsson B, Hampson G, Michaels J, et al. Advanced therapy medicinal products and health technology assessment principles and practices for value-based and sustainable healthcare. Eur J Health Econ. 2019;20:427–38.
15. Drummond MF, Neumann PJ, Sullivan SD, et al. Analytic considerations in applying a general economic evaluation reference case to gene therapy. Value Health. 2019;22:661–8.
16. European Network for Health Technology Assessment. Diemen, The Netherlands. https://eunethta.eu/wp-content/uploads/2018/01/2016-03-07_reflection_paper_pm_2nd_draft.pdf. Accessed 27 June 2019.
17. Office of Health Economics. London, UK. https://www.ohe.org/publications/exploring-assessment-and-appraisal-regenerative-medicines-and-celltherapy-products. Accessed 22 Mar 2019.
18. Rogowski W, Payne K, Schnell-Inderst P, et al. Concepts of 'personalization' in personalized medicine: implications for economic evaluation. Pharmacoeconomics. 2015;33:49–59.
19. Love-Koh J, Peel A, Rejon-Parrilla JC, et al. The future of precision medicine: potential impacts for health technology assessment. Pharmacoeconomics. 2018;36:1439–51.
20. Lakdawalla DN, Doshi JA, Garrison LP Jr, et al. Defining elements of value in health care: a health economics approach: an ISPOR Special Task Force report [3]. Value Health. 2018;21:131–9.
21. Bramer WM, Giustini D, de Jonge GB, et al. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016;104:240–3.
22. Othus M, Barlogie B, Leblanc ML, et al. Cure models as a useful statistical tool for analyzing survival. Clin Cancer Res. 2012;18:3731–6.
23. National Institute for Health and Care Excellence. London, UK. https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technologyappraisal-2013-pdf-2007975843781. Accessed 26 Mar 2020.
24. Garrison LP, Towse A. Value-based pricing and reimbursement in personalised healthcare: introduction to the basic health economics. J Pers Med. 2017;7:10.
25. Garrison LP Jr, Zamora B, Li M, et al. Augmenting cost-effectiveness analysis for uncertainty: the implications for value assessment: rationale and empirical support. J Manag Care Spec Pharm. 2020;26:400–6.
26. Attema AE, Brouwer WB, Claxton K. Discounting in economic evaluations. Pharmacoeconomics. 2018;36:745–58.
27. Kip MM, IJzerman MJ, Henriksson M, et al. Toward alignment in the reporting of economic evaluations of diagnostic tests and biomarkers: the AGREEDT checklist. Med Decis Mak. 2018;38:778–88.

28. US Food and Drugs Administration. Silver Spring, Maryland, US. https://www.fda.gov/drugs/fda-expands-pembrolizumab-indication-first-linetreatment-nsclc-tps-1. Accessed 14 Sept 2020.

29. Garrison LP Jr, Lalla D, Brammer M, *et al*. Assessing the potential cost-effectiveness of retesting IHC0, IHC1+, or FISH-negative early stage breast cancer patients for HER2 status. Cancer. 2013;119:3113–22.

30. Whittington MD, McQueen RB, Ollendorf DA, *et al*. Long-term survival and value of chimeric antigen receptor T-cell therapy for pediatric patients with relapsed or refractory leukemia. JAMA Pediatr. 2018;172:1161–8.

31. Park JJH, Hsu G, Siden EG, *et al*. An overview of precision oncology basket and umbrella trials for clinicians. CA Cancer J Clin. 2020;70:125–37.

32. Cunanan KM, Iasonos A, Shen R, *et al*. An efficient basket trial design. Stat Med. 2017;36:1568–79.

33. Percha B, Baskerville EB, Johnson M, *et al*. Designing robust n-of-1 studies for precision medicine: simulation study and design recommendations. J Med Internet Res. 2019;21:e12641.

34. Pallmann P, Bedding AW, Choodari-Oskooei B, *et al*. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16:29.

35. Hatswell AJ, Baio G, Berlin JA, *et al*. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. BMJ Open. 2016;6:e011666.

36. Goring S, Taylor A, Muller K, *et al*. Characteristics of non-randomised studies using comparisons with external controls submitted for regulatory approval in the USA and Europe: a systematic review. BMJ Open. 2019;9:e024895.

37. Grieve R, Abrams K, Claxton K, *et al*. Cancer Drugs Fund requires further reform. BMJ. 2016;354:i5090.

38. Van De Wetering E, van Exel J, Brouwer WB. The challenge of conditional reimbursement: stopping reimbursement can be more difficult than not starting in the first place! Value Health. 2017;20:118–25.

39. McCabe C, Chilcott J, Claxton K, *et al*. Continuing the multiple sclerosis risk sharing scheme is unjustified. BMJ. 2010;340:c1786.

40. Garralda E, Dienstmann R, Piris-Gimenez A, *et al*. New clinical trial designs in the era of precision medicine. Mol Oncol. 2019;13:549–57.

41. Gyawali B, Hey SP, Kesselheim AS. Assessment of the clinical benefit of cancer drugs receiving accelerated approval. JAMA Intern Med. 2019;179:906–13.

42. Bazhenova L, Jiao X, Lokker A, et al. Cancers with NTRK gene fusions: molecular characteristics and prognosis. American Association for Cancer Research Annual Meeting; 22–24 June, 2020; virtual.

43. Heramb C, Wangensteen T, Grindedal EM, *et al*. BRCA1 and BRCA2 mutation spectrum: an update on mutation distribution in a large cancer genetics clinic in Norway. Hered Cancer Clin Pract. 2018;16:3.

44. Wiesch DG, Meyers DA. Strategies for analyzing genotype-phenotype relationships in asthma. J Allergy Clin Immunol. 2000;105:S482–6.

45. Wordsworth S, Doble B, Payne K, *et al*. Using "big data" in the cost-effectiveness analysis of next-generation sequencing technologies: challenges and potential solutions. Value Health. 2018;21:1048–53.

46. Kazi DS, Garber AM, Shah RU, *et al*. Cost-effectiveness of genotype-guided and dual antiplatelet therapies in acute coronary syndrome. Ann Intern Med. 2014;160:221–32.

47. Eggemann H, Ignatov T, Burger E, *et al*. Moderate HER2 expression as a prognostic factor in hormone receptor positive breast cancer. Endocr Relat Cancer. 2015;22:725–33.

48. Ouwens MJ, Mukhopadhyay P, Zhang Y, *et al*. Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. Pharmacoeconomics. 2019;37:1129–38.

49. Janssen-Heijnen M, Houterman S, Lemmens V, *et al*. Prognosis for long-term survivors of cancer. Ann Oncol. 2007;18:1408–13.

50. Institute for Clinical and Economic Review. Boston, Massachusetts, US. https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101744954-pdf. Accessed 22 Mar 2019.

51. Foster MW, Mulvihill JJ, Sharp RR. Evaluating the utility of personal genomic information. Genet Med. 2009;11:570–4.

52. Regier DA, Weymann D, Buchanan J, *et al*. Valuation of health and nonhealth outcomes from next-generation sequencing: approaches, challenges, and solutions. Value Health. 2018;21:1043–7.

53. Lee DW, Neumann PJ, Rizzo JA. Understanding the medical and nonmedical value of diagnostic testing. Value Health. 2010;13:310–4.

54. Lakdawalla DN, Romley JA, Sanchez Y, *et al*. How cancer patients value hope and the implications for cost-effectiveness assessments of high-cost cancer therapies. Health Aff (Millwood). 2012;31:676–82.

55. Thokala P, Ochalek J, Leech AA, *et al*. Cost-effectiveness thresholds: the past, the present and the future. Pharmacoeconomics. 2018;36:509–22.

56. Asphaug L, Melberg HO. The cost-effectiveness of multigene panel testing for hereditary breast and ovarian cancer in Norway. MDM Policy Pract. 2019;4:2381468318821103.

57. Dias LC, Morton A, Quigley J. Elicitation, vol. 1. New York: Springer International Publishing; 2018. p. 3.

58. Bojke L, Claxton K, Sculpher M, *et al*. Characterizing structural uncertainty in decision analytic models: a review and application of methods. Value Health. 2009;12:739–49.

59. Grigore B, Peters J, Hyde C, *et al*. EXPLICIT: a feasibility study of remote expert elicitation in health technology assessment. BMC Med Inform Decis Mak. 2017;17:131.

60. Knol AB, Slottje P, van der Sluijs JP, *et al*. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environ Health. 2010;9:19.

61. Bojke L, Claxton K, Bravo-Vergel Y, *et al*. Eliciting distributions to populate decision analytic models. Value Health. 2010;13:557–64.

62. Bilcke J, Beutels P, Brisson M, *et al*. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. Med Decis Mak. 2011;31:675–92.

63. Antonanzas F, Juárez-Castelló C, Lorente R, Rodríguez-Ibeas R. The use of risk-sharing contracts in healthcare: theoretical and empirical assessments. Pharmacoeconomics. 2019;37:1469–83.

64. Lin JK, Lerman BJ, Barnes JI, *et al*. Cost effectiveness of chimeric antigen receptor T-cell therapy in relapsed or refractory pediatric B-cell acute lymphoblastic leukemia. J Clin Oncol. 2018;36:3192–202.

65. Philips Z, Bojke L, Sculpher M, *et al*. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics. 2006;24:355–71.

66. Evers S, Goossens M, de Vet H, *et al*. Criteria list for assessment of methodological quality of economic evaluations: consensus on health economic criteria. Int J Technol Assess Health Care. 2005;21:240–5.

67. Husereau D, Drummond M, Petrou S, *et al*. Consolidated Health Economic Evaluation Reporting Standards (CHEERS): explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health. 2013;16:231–50.

68. Chiou CF, Hay JW, Wallace JF, *et al*. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care. 2003;41:32–44.

69. Vemer P, Corro Ramos I, van Voorn GA, *et al*. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. Pharmacoeconomics. 2016;34:349–61.

70. Huls SPI, Whichello CL, van Exel J, *et al*. What is next for patient preferences in health technology assessment? A systematic review of the challenges. Value Health. 2019;22:1318–28.

71. Annemans L, Redekop K, Payne K. Current methodological issues in the economic assessment of personalized medicine. Value Health. 2013;16(6 Suppl.):S20–6.

72. Payne K, Gavan SP, Wright SJ, *et al*. Cost-effectiveness analyses of genetic and genomic diagnostic tests. Nat Rev Genet. 2018;19:235–46.

73. Phillips KA, Deverka PA, Marshall DA, *et al*. Methodological issues in assessing the economic value of next-generation sequencing tests: many challenges and not enough solutions. Value Health. 2018;21:1033–42.

74. Yang Y, Abel L, Buchanan J, *et al*. Use of decision modelling in economic evaluations of diagnostic tests: an appraisal and review of health technology assessments in the UK. Pharmacoecon Open. 2019;3:281–91.

75. Degeling K, Koffijberg H, IJzerman MJ. A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. Expert Rev Pharmacoecon Outcomes Res. 2017;17:17–25.

76. Christensen KD, Phillips KA, Green RC, *et al*. Cost analyses of genomic sequencing: lessons learned from the MedSeq Project. Value Health. 2018;21:1054–61.

77. Marshall DA, Grazziotin LR, Regier DA, *et al*. Addressing challenges of economic evaluation in precision medicine using dynamic simulation modeling. Value Health. 2020;23:566–73.

78. Corro Ramos I, Hoogendoorn M, Rutten-van Mölken M. How to address uncertainty in health economic discrete-event simulation models: an illustration for chronic obstructive pulmonary disease. Med Decis Mak. 2020;40:619–32.

79. Bennette CS, Gallego CJ, Burke W, *et al*. The cost-effectiveness of returning incidental findings from next-generation genomic sequencing. Gen Med. 2015;17:587–95.

80. Thokala P, Devlin N, Marsh K, *et al*. Multiple criteria decision analysis for health care decision making: an introduction. Report 1 of the ISPOR MCDA Emerging Good Practices Task Force. Value Health. 2016;19:1–13.