Validating a Budget Impact Model Using Payer Insight and Claims Data: A Framework and Case Study

Anna Hung, PharmD, PhD, MS; Julia F. Slejko, PhD; Amy Lugo, PharmD; Fadia Shaya, PhD; Stuart T. Haines, PharmD; and C. Daniel Mullins, PhD

ABSTRACT

BACKGROUND: There is a paucity of studies validating budget impact models. The lack of such studies may contribute to the underuse of budget impact models by payers in formulary decision making.

OBJECTIVE: To assess the face validity, internal verification, and predictive validity of a previously published model that assessed the budgetary impact of antidiabetic formulary changes.

METHODS: 4 experts with diverse backgrounds were selected and asked questions regarding the face validity of the structure/conceptual model, input data, and results from the budget impact model. To assess internal verification, structured “walk-throughs,” unit tests, extreme condition tests, traces, replication tests, and double programming techniques were used. The predictive validity of the model was evaluated by comparing the predicted and realized budget using mean absolute scaled error. “Realized” budgetary impact of the formulary changes was calculated by taking the difference between realized budget in the year after the formulary changes and the budget had there been no formulary changes (i.e., the counterfactual). The counterfactual budget was modeled using the best fit autoregressive integrated moving average model.

RESULTS: When assessing the face validity of the model, the 4 experts brought up issues such as how to incorporate other health insurance, recent policy changes, cost inflation, and potential impacts on insulin use. The 6 internal verification techniques caught mistakes in equations, missing data, and misclassified data. The realized budget was found to be lower than the predicted budget, with 13% error and an absolute scaled error of 2.60. After removing the model assumption that past utilization trends would continue, the model's predictive accuracy improved (the absolute scaled error dropped below 1 to 0.48). The “realized” budgetary impact was found to be greater than the predicted budgetary impact, largely because of lower-than-expected utilization.

CONCLUSIONS: The budget impact model overpredicted utilization in the year after the formulary changes. Discoveries through the validation process improved the accuracy and transparency of the model.

J Manag Care Spec Pharm. 2019;25(9):913-21

Copyright© 2019, Academy of Managed Care Pharmacy. All rights reserved.

What this study adds

- This study assessed the face validity, internal verification, and predictive validity of a previously published budget impact model.
- This study demonstrated how undergoing a validation process for a budget impact model can lead to improved accuracy and credibility.
- This study serves as an example for future health plans, managed care organizations, sponsors of budget impact models, and budget impact model researchers on how to undergo and document a validation process for a budget impact model.

Validation is expected to establish credibility for models. Part of the reason sponsor-provided budget impact models are rarely used by payers in decision making is the lack of validation. While there has been recent interest in validating budget impact models, there is a paucity of published studies that validate the predictive validity of a budget impact model.¹⁻⁴

The 2014 International Society for Pharmacoeconomics and Outcomes Research (ISPOR) guidelines for budget impact analysis (BIA) mention 3 types of validation for a budget impact model: (a) face validity, (b) internal verification, and (c) predictive validity.⁵

Part 1. Face Validity

Face validity is defined as “the extent to which a model, its assumptions, and applications correspond to current science and evidence, as judged by people who have expertise in the problem.”⁶ In a broader sense, a model has face validity if experts believe the model and its behavior are reasonable.⁷ If aspects of the model do not make intuitive sense, these discrepancies and their potential impact on results should be explained.^{6,8} Face validity should be assessed by experts who are impartial to the results of the analysis.⁹

There are 3 sets of guidelines for model validation that discuss face validity and can be applied to budget impact models: the 2012 ISPOR-Society for Medical Decision Making (SMDM),⁶ 2014 ISPOR-Academy of Managed Care Pharmacy (AMCP)–National Pharmaceutical Council (NPC),¹⁰ and 2015 Assessment of the Validation Status of Health Economic decision models (AdVisHE).¹¹ These agree that face validity

What is already known about this subject

- Budget impact models predict the financial consequences of a new drug or technology entering the market; however, the use of budget impact models by U.S. payers to inform drug formulary decision making is low.
- Validation increases the credibility of a model, increasing the probability of use.

should be assessed in 3 separate areas related to the model: (a) structure/conceptual model, (b) data, and (c) results. Based on this guidance, as well as ISPOR BIA guidelines, we developed a 2-part questionnaire to select and ask experts when assessing face validity for budget impact models (Appendix A, available in online article).^{5,6,10-12}

Part 2. Internal Verification

Internal verification addresses whether the mathematical calculations and computer programming of the model have been performed correctly and are consistent with the model's specifications.^{6-8,10} Internal verification is also known as verification, internal validity, internal consistency, internal testing, technical validity, technical validation, computerized model validity, and debugging.^{6,8,10,11,13} Interestingly, when a Delphi panel was conducted, internal validity was interpreted differently by several members.¹¹

There are 3 guidelines on model validation that discuss internal verification and can be applied to budget impact models: the 2012 ISPOR-SMDM,⁶ 2014 ISPOR-AMCP-NPC,¹⁰ and 2015 AdVisHE.¹¹ The 2 main steps of internal verification include: (a) verifying individual equations and input parameters against their sources and (b) checking that code was accurately implemented.^{6,10} This should be done systematically.¹⁰

The following are examples of techniques that can be used to assess internal verification.

- *Structured "walk-throughs":* Explaining the code to others who are searching for errors⁶
- *Unit tests:* Examining separate parts of a model one by one^{6,11}
- *Extreme condition tests:* Inputting extremely high or low input values and assessing if the model outputs are plausible^{6,7,11,13}
- *Traces:* Following (or tracing) units (often patients) through the model and assessing if the model logic is behaving appropriately^{6,7,11}
- *Replication tests:* Running the program multiple times with equivalent input values and determining if model outputs are the same¹³
- *Double programming:* Independent programming by 2 different people or using 2 types of software⁶

Part 3. Predictive Validity

Predictive validity relates to how closely a model forecasts observed events in the future.⁶ Predictive validity is also known as predictive accuracy and forecast accuracy.¹⁴ Predictive validity is the strongest form of validation and should be performed whenever feasible because predictive validation reveals how well a model achieved its intended purpose.^{6,10} Predictive validation is compelling because it is an objective assessment and cannot be influenced to fit the observed data in some way.^{6,10} However, a model should not be criticized for poor predictive validity because unforeseen circumstances may occur in the

future.^{13,15} Instead, a model should be recalibrated to adapt to new evidence as it becomes available.¹³

Although model validation guidance, such as the 2012 ISPOR-SMDM and 2014 ISPOR-AMCP-NPC,^{6,10} mention conducting predictive validation, they do not provide in-depth suggestions regarding the types of forecast error measures to use when comparing model forecasts to actual observed events. Common forecast error measures include R^2 , mean absolute error, mean absolute percentage error, and root mean squared error.^{14,16-20} Other measures include absolute difference, median absolute error, median absolute percentage error, and mean absolute scaled error.²¹

Hyndman et al. (2006) discuss the advantages and disadvantages of different types of measures.²¹ The absolute error measures are easy to understand but are scale-dependent. Thus, they should only be used when scales are the same. The absolute percentage error measures are scale-independent; however, if the actual value, which is in the denominator, is very small and approaches zero, the measure becomes very large and approaches infinity. The relative error measures are calculated as a ratio of the error from the current forecasting method and the error from a naive forecasting method, often the random walk method. Relative error measures also may become very large and approach infinity if the error from the random walk method, which is in the denominator, is very small and approaches zero (e.g., when the last value is close to or the same as the actual value). Hyndman et al. ultimately recommend using the mean absolute scaled error (MASE) measure. This measure is scale-independent and does not have issues of approaching infinity, unless every single actual value is the same as the last value, which is rare. MASE is calculated as:

MASE = mean($|q_t|$), where

$$q_t = \frac{e_t}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|}$$

When MASE is less than 1, the forecast performs better than the average 1-step naive forecast computed in-sample. When determining whether a model is "good," demand forecasting literature advises against setting an arbitrary threshold such as mean absolute percentage error less than 20%.²² Instead, at a minimum, a model will add value if it performs better than assuming next year's expenditures will be the same as last year's expenditures (unless no budget impact is expected), which over time is equivalent to when MASE is less than 1.

Case Study: Antidiabetic Formulary Management for TRICARE

A previously published model predicted the budgetary impact of antidiabetic formulary changes for the TRICARE payer.²³ The formulary changes of interest involved preferred agent policies, step therapy requirements, and an addition of a drug

on one of the formularies (see Appendix B, available in online article). The objective of this study was to test the face validity, internal verification, and predictive validity of the previously published budget impact model.

Methods

The previously published budget impact model was developed according to ISPOR BIA guidelines to help inform payer formulary decision making (see Figure 2 from the original budget impact model publication).²³ The Excel-based model used a 3-year time horizon, payer perspective, and health plan claims data to estimate model inputs. Key model assumptions are provided in Appendix A in the original budget impact model publication.²³

Face validity and internal verification, which could be assessed before the implementation of the formulary changes, were assessed before the budget impact model presented in the previous publication.²³ However, predictive validity, which could not be assessed until after the implementation of the formulary changes, was not assessed until after the budget impact model in the previous publication had been presented.

Part 1. Face Validity

Four experts (3 faculty members and 1 individual employed by the payer) were selected to review the budget impact model and raise any issues after considering the Budget Impact Model Face Validity Questionnaire (Appendix A). The experts' recommendations and responses to these experts were documented.

Part 2. Internal Verification

Structured "walk-throughs," unit tests, extreme condition tests, traces, replication tests, and double programming were used during the internal verification process. For structured walk-throughs, experts helping assess face validity of the model were walked through each of the main input and output tabs in the Microsoft Excel workbook (Microsoft Corp., Redmond, WA). For unit tests, the modeler checked that everything was working correctly for each tab, and specifically, the modeler compared input data to the source and made sure individual equations were accurate.

Extreme condition tests were based on assessing results from sensitivity analyses and subgroup analyses. Some of these analyses provided good test cases of very little utilization. Traces were conducted by looking at key model outputs and making sure they were correct by tracing their calculations back to the original data. Traces were also commonly used alongside other techniques. When another technique led to questionable results, these model outputs were traced back to the problem.

For replication tests, the modeler verified that the results did not change when rerunning all macros. Double programming

typically involves programming by another person or in another type of software. In our case, the original modeler reprogrammed sections of tabs in SAS version 9.3 (SAS Institute, Cary, NC) to see if the same results were seen.

Part 3. Predictive Validity

The original budget impact model estimated the budget and the budgetary impact for the year after the initial implementation date of the antidiabetic formulary changes. These estimates were compared with realized costs in the year after implementation.

Year 1 Budget: Comparing Actual Versus Predicted. As described in the previous publication, the original budget impact model used health plan claims data from before the implementation of the formulary changes to determine the baseline budget for the utilization of antidiabetic drugs and drugs used to treat side effects from the antidiabetic drugs in the base-case scenario.²³ This baseline budget was then used to estimate the future budget with the implementation of the formulary changes. One year after the formulary changes of interest, we used newly available health plan claims data and the same methods to determine the realized budget from the year after the formulary changes. As in the original model, the budget was subdivided into utilization costs, rebates, copays, and dispensing fees. Because actual costs could not be published, this study used pharmacy and medical cost estimates from publicly available sources.²³ While unit costs were the same as in the model, the weighted average cost of a 30-day fill for a given drug could change if the utilization mix of National Drug Code numbers for a given drug changed or if the quantity for a given day supply changed. The difference was calculated as realized budget minus predicted budget. Percentage differences were calculated as the difference divided by the absolute value of the realized budget. Mean and median percentage and absolute percentage differences across different cost categories, such as utilization costs, rebates, copays, and dispensing fees across the antidiabetic drugs as well as drugs used to treat side effects, were also reported.

Year 1 Budgetary Impact. The "realized" budgetary impact of these formulary changes was calculated as the difference between the budget had the formulary changes not occurred (i.e., the counterfactual) and the realized budget after the formulary changes. To calculate the counterfactual budget, we estimated the budget assuming no formulary change using autoregressive integrated moving average models and the Box-Jenkins framework.²⁴ The final model was chosen based on (a) the lowest Akaike information criterion and Schwarz Bayesian criteria if models were of the same type (e.g., not comparing seasonal vs. nonseasonal models or differenced vs. nondifferenced models); (b) mean absolute percentage error and root mean squared error values if models were not of the same type;

TABLE 1 Summary of Key Recommendations from Face Validity Check

Topic	Recommendation from Expert	Response to Recommendation
Structure/conceptual model	<ul style="list-style-type: none"> • Add Soliqua (insulin glargine and lixisenatide injection, Sanofi-Aventis) • Decrease market share shift of Byetta (exenatide, AstraZeneca) to exenatide (generic) from 5% to 1% • Consider adjusting prices for inflation in the future • Need to lower costs paid for prescriptions for those with OHI 	<ul style="list-style-type: none"> • Added Soliqua • Decreased market share shift of Byetta to exenatide (generic) from 5% to 1% • Considered adjusting prices but discussed with a budget impact modeling expert and since it is difficult to justify an estimate for future inflation, we did not include. Listed as limitation instead • Incorporated OHI into the model. It was assumed that for those with OHI, the payer paid 30% of pharmacy costs
Data	Keep in mind that mandatory mail and MTF policies for specific branded maintenance drugs went into effect March 2014 and October 2015. Thus, the “natural” utilization trend from February 2014-January 2015 to February 2015-January 2016 may incorporate these utilization shifts from retail to mail or MTF point-of-service and we would not expect these shifts in our study time period	This was listed as a limitation
Results	Consider adding a sensitivity analysis scenario that basal insulin could be used more because of decreased access to GLP-1RA (10% decrease to 20% increase in insulin use)	Considered adding this as a sensitivity analysis but after discussing further with a payer expert, we decided not to because it is too easy for prescribers to get around the PA criteria and the insulins have their own preferred products
Other	<ul style="list-style-type: none"> • Change location of legend on input tab 4 of model • Have output tab 1 of model results pop up after each sensitivity analysis • There are glitches when sensitivity analyses run (tables flash and disappear) • On sensitivity analysis tab, add notes to describe each tab 	<ul style="list-style-type: none"> • Changed location of legend • Set macros so that output tab 1 results pop up after each sensitivity analysis • Removed extra objects so fewer glitches when sensitivity analyses run • Added notes on sensitivity analysis tab

GLP-1RA=glucagon-like peptide-1 receptor agonist; MTF=military treatment facility; OHI=other health insurance; PA=prior authorization.

(c) visual checks to make sure there were no major deviations of the forecast from the data; and (d) *P* values greater than 0.05 in the Ljung-Box white noise residual test.

These analyses were conducted in SAS version 9.3. This study received approval from the University of Maryland Baltimore and Defense Health Agency institutional review boards.

Results

Part 1. Face Validity

The 4 experts selected were knowledgeable in different areas: the payer formulary changes of interest, diabetes treatment, modeling, and payer claims data. For the most part, the model seemed to have face validity. However, there were several issues that were brought up and resolved (Table 1). For example, in the model, the payer did not pay less for those with other health insurance for their pharmacy benefit. Thus, this was later built into the model with an assumption that the payer would pay 30% of pharmacy costs for those with other health insurance for the pharmacy benefit.

Another key issue that was identified was that other policy changes affecting where patients received their medications had occurred in the baseline year or year before the formulary change. This meant that only 2 months of the baseline year

data would be reflective of future utilization without the formulary change when considering where patients received their medications. Given that this affected mostly where patients received their medications and not necessarily the formulary changes of interest, this issue was added to the list of limitations for the model.

A third issue that was raised was whether costs should be adjusted for inflation in the future. ISPOR BIA guidelines recommend that an attempt be made to forecast these if the assumptions can be justified and supported by evidence if feasible.⁵ Considering the short time horizon, that there is not much evidence for the future, and predicting inflation without such evidence could introduce bias, we did not adjust for inflation of drug and medical costs. Instead, this was noted in the model's limitations section.

A final issue that was identified was whether the selection of a preferred glucagon-like peptide-1 receptor agonist (GLP-1RA) would lead to decreased utilization of basal insulin. Considering that the payer GLP-1RA prior authorization criteria could be easily satisfied and the basal insulins could have their own preferred products, we decided to keep the expectation that a preferred GLP-1RA would not significantly decrease utilization of basal insulin.

TABLE 2 Key Findings and Updates from Internal Verification Process

Technique	Implementation	Findings and Changes
Structured “walk-throughs”	Went through main input and output tabs in the model spreadsheet with each face validity expert	Findings, such as wrong market share percentages in the input tab 3 of the model, were reported in the face validity section
Unit tests	Looked through each tab, checked equations, compared data with input data, and searched for errors	Found and corrected mistakes in input data where values were out of order, missing, or misclassified Found and corrected errors in equations that did not account for OHI versus non-OHI after recommendations from the face validity process Changes to improve formatting and readability
Extreme condition tests: sensitivity and subgroup analyses	Examined results for each sensitivity and subgroup analysis. Some of these sensitivity and subgroup analyses had very little utilization Created additional scenarios combining sensitivity analyses, subgroup analyses, and artificial data. Compared these results to what would be expected	Found odd trends over time in budget and budget impact. Traced these back to errors in equations; fixed these errors Found DIV/0 errors in the output tab. Traced these back to the lack of utilization of a drug class for some subset models. Fixed
Traces	Traced back key budget and budget impact results on output tabs When errors were found, we traced errors back to the original input data	See above
Replication tests	Checked that running all macros (e.g., reset buttons, sensitivity analysis buttons, cost driver buttons) would lead to the same results	No errors found
Double programming	Reprogrammed sections of tabs and checked results with original tab Reprogrammed calculation of the baseline year’s net budget in SAS and found errors in data input values on copay tabs	Corrected data input values on copay tabs

DIV/0 = a division by 0 error that occurs in Microsoft Excel; OHI = other health insurance.

Part 2. Internal Verification

Table 2 shows some of the key findings from the internal verification process. The findings included mistakes in equations; input data that were missing, out of order, or misclassified; DIV/0 errors (a division by 0 error that occurs in Microsoft Excel); and improvement in formatting or readability. After other modifications were made to the model based on recommendations from the face validity process, new errors related to these modifications surfaced. These errors were fixed before reporting of the model’s predictions.

Part 3. Predictive Validity

Year 1 Budget. Table 3A shows that the realized budget was generally lower than the predicted budget for Year 1 (mean of 31% lower and median of 21% lower). Overall, the realized net budget was \$79 million lower (13% error) for antidiabetic drugs directly affected by the formulary changes and \$5 million (67% error) lower for drug side effect treatment than what was predicted.

Among the cost categories, the largest percentage difference seen for spending on antidiabetic drugs was the dispensing fee (28% less), followed by copay offset (25% less). The largest percentage difference seen for drug side effect treatment was for rebates (63% less), followed by utilization (54% less).

The absolute scaled error for the net budget for antidiabetic drugs was 2.60 (Table 3A), which indicated that the model performed worse than an average 1-step naive forecast computed in-sample (e.g., last year’s figures). One major reason for the poor forecasting accuracy was the model assumption that any increases or decreases 2 years to 1 year before the formulary change would continue forward. When this assumption was removed (e.g., no trend was assumed), the absolute scaled error for the net budget for antidiabetic drugs directly affected by the formulary changes decreased to 0.48 (Table 3B).

Year 1 Budgetary Impact. Based on lowest mean absolute percentage error and root mean squared error values, a log model with an autoregressive term of 3 was chosen as the final model for the budget had no formulary changes been implemented (i.e., the counterfactual). This model predicted \$657 million in spending from February 2016 to January 2017 had there been no formulary changes. This was compared with the actual \$609 million in spending for the antidiabetic drugs that was seen from February 2016 to January 2017. This resulted in a budgetary impact of \$49 million (95% confidence interval: \$6 million to \$91 million) in the year after implementation of the formulary changes.

TABLE 3 Forecast Accuracy of Model When Predicting Budget in Year After Formulary Change

	YIFC Predicted, \$	YIFC Empirical, \$	Difference (Empirical–Predicted), \$	Percentage Error	Absolute Scaled Error
A. Forecast Accuracy of YIFC Budget					
AntiDM_dir					
Utilization	815,391,798	716,121,489	-99,270,309	-14	2.51
-Rebate	-102,226,967	-86,815,930	15,411,037	-18	2.93
-Copay	-28,198,789	-22,559,024	5,639,765	-25	1.25
+Dispensing fee	2,496,055	1,948,811	-547,243	-28	1.12
=Net budget	687,462,097	608,695,347	-78,766,750	-13	2.60
Side effect					
Utilization	19,562,776	12,729,267	-6,833,509	-54	1.81
-Rebate	-3,970,627	-2,440,100	1,530,528	-63	1.83
-Copay	-3,987,568	-3,551,600	435,968	-12	0.91
+Dispensing fee	713,334	619,841	-93,494	-15	2.31
=Net budget	12,317,916	7,357,408	-4,960,507	-67	1.44
Mean				-31	1.87
Median				-21	1.82
B. Forecast Accuracy of YIFC Budget When Assuming No Trend					
AntiDM_dir					
Utilization	742,100,535	716,121,489	-25,979,045	-4	0.66
-Rebate	-93,619,827	-86,815,930	6,803,897	-8	1.30
-Copay	-27,767,038	-22,559,024	5,208,014	-23	1.15
+Dispensing fee	2,443,241	1,948,811	-494,429	-25	1.01
=Net budget	623,156,911	608,695,347	-14,461,564	-2	0.48
Side effect					
Utilization	16,495,704	12,729,267	-3,766,437	-30	1.00
-Rebate	-3,276,836	-2,440,100	836,736	-34	1.00
-Copay	-3,826,162	-3,551,600	274,562	-8	0.57
+Dispensing Fee	659,956	619,841	-40,116	-6	0.99
=Net budget	10,052,662	7,357,408	-2,695,254	-37	0.78
Mean				-18	0.89
Median				-15	0.99

AntiDM_dir = antidiabetic drugs directly affected by formulary changes; YIFC = Year 1 formulary change.

As seen in Figure 1, the final model for no formulary change consistently forecasted spending greater than actual spending after February 2016.

There was a substantial difference in the budgetary impact predicted in the original model, \$24 million, and the \$49 million “realized” in the year after the formulary changes. This was partly because of lower-than-predicted utilization, again related to the assumption that past utilization trends would continue forward.

Discussion

Previous studies have found lower than expected use of published budget impact models by payers, citing barriers such as perceived bias, lack of transparency and understanding, and model complexity.^{25–28} Independent researchers, as well as organizations such as ISPOR, AMCP, and NPC, have put forth guidance on validation methods to assess the relevance and credibility of a model.^{6,7,9–11,13,29} However, none has been specific

to budget impact models. ISPOR BIA guidelines mention the importance of validation but do not provide examples.^{5,12}

Few examples of budget impact model validation are available in published literature. Two budget impact analyses that mentioned validation used discrete event simulation and were focused on demonstrating that the simulation model was valid, such as prediction of prevalence and incidence, as opposed to the budget impact.^{1,2} For both of these articles, validation was mentioned as a small part of the article.

One poster presented at the ISPOR Annual International Meeting in 2015 had a primary objective of validating a budget impact model for the use of denosumab in postmenopausal women with osteoporosis.³ The authors evaluated the face validity of the model by comparing the underlying Markov model with prior published cost-effectiveness studies. Next, the authors evaluated the internal validity of the model using extreme value input parameters and 2-way sensitivity analyses varying market share, drug price, and direct medical costs.

FIGURE 1 Counterfactual Budget (Had No Formulary Changes Occurred) Versus Actual Budget in the Year After Formulary Changes

FC=formulary change; LOG_A3=logarithmic model with autoregressive term of 3.

Another poster presented at the AMCP Annual Meeting in 2017 had a primary objective of validating a budget impact model for a clinical management program for a biosimilar.⁴ This study assessed predictive validity by updating model inputs with actual switch rates from a nonpreferred to preferred drug, actual utilization, actual drug costs, as well as modifying the model structure to accommodate unforeseen switching to a third drug.

The methods presented in this study provide a framework for future researchers to validate their budget impact models. Specific tools, including a face validity questionnaire, an internal verification template, and a review of forecasting accuracy measures for predictive validity are presented. A previously published budget impact model predicting the financial consequences of a payer's antidiabetic medication formulary changes was used as a case study of how to apply the 3 types of validation and tools to budget impact models.²³

Throughout the face validity process, the model was updated as experts from diverse areas reviewed and ensured significant components were represented and linked as expected in the model. Next, the internal verification process was necessary to ensure the integrity of the model. Finally, comparing actual budget expenditures versus what was predicted (i.e., the

predictive accuracy of the model) gave a sense of how well the model performed. We explored what was leading to overestimates of budget expenditures, and it was not until model assumptions were altered that our model provided better estimates. Hence, a systematic validation process led to an improved and more credible model in our case, and is a critical step to improving the accuracy of a budget impact model. In an ideal scenario, a payer could continuously evaluate the predictive validity of a budget impact model as more time passes and new claims data become available. Then, a payer could identify whether the budget impact model's predictive validity held true over time, thereby increasing confidence in the model's predictions.

One of the strengths of this validation process was collaborating with the end user (e.g., payer) of the model. This allowed for a payer representative to serve as an expert during the face validation process and ensure that key components were included. A key advantage of payer collaboration is having access to payer claims data to be able to check the predictive validity of the model, as well as to inform model inputs. Altogether, a model created in collaboration with the payer resulted in a model that was more usable to the payer.

In addition to face validity, internal verification, and predictive validity, other types of validation could be useful. Between-model validation can be helpful to validate budget impact models if another model exists. However, this is not often the case. An exploration and comparison of cost drivers between the forecasted and empirical analysis can also aid the validation and recalibration processes.

Limitations

This validation study has some limitations to consider. First, there is a lack of real pharmacy and medical costs. Since this information was considered proprietary, publicly available unit costs of medications and medical visits were used instead. However, a validation study reported internally to payers could incorporate actual payer costs. A second limitation is the inability to see the actual budget had the formulary changes not occurred (e.g., the counterfactual). However, there are ethical and operational barriers to running a randomized trial that imposes formulary changes to some but not others in a real-world payer setting.

Conclusions

This is one of the few budget impact model validation case studies presented in the literature. This article discussed and then demonstrated how to assess face validity, internal verification, and predictive validity of a budget impact model. Through our validation process that included end-user collaboration, we improved the model through corrections, as well as modified a key assumption.

Authors

ANNA HUNG, PharmD, PhD, MS, Duke Clinical Research Institute, Durham, North Carolina. JULIA F. SLEJKO, PhD; FADIA SHAYA, PhD; and C. DANIEL MULLINS, PhD, University of Maryland School of Pharmacy, Baltimore. AMY LUGO, PharmD, Defense Health Agency, San Antonio, Texas, and STUART T. HAINES, PharmD, the University of Mississippi School of Pharmacy, Jackson.

AUTHOR CORRESPONDENCE: Anna Hung, PharmD, PhD, MS, Duke Clinical Research Institute, 200 Morris St., Durham, NC 27713. Tel.: 301.412.7396; E-mail: anna.hung@duke.edu.

DISCLOSURES

This project was supported by grant number F32HS024857 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent the official views of AHRQ. AHRQ had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or design to submit the manuscript for publication. The findings discussed in this manuscript represent the views of the authors and do not necessarily reflect the views of the Department of Defense, the Defense Health Agency, nor the Departments of the Army, Navy, and Air Force.

Hung reports a grant from the AHRQ, during the conduct of the study, and personal fees from CVS Health and BlueCross BlueShield Association, outside the submitted work. Mullins reports grants and personal fees from Bayer and Pfizer and personal fees from Boehringer Ingelheim, Janssen/J&J, Regeneron, and Sanofi-Aventis, outside the submitted work. Mullins, Slejko, and Shaya are employed by the University of Maryland School of Pharmacy. Haines and Lugo have nothing to disclose.

Part of this content was previously presented as a poster at the 2017 AMCP Managed Care & Specialty Pharmacy Annual Meeting; March 27-30, 2017; Denver, CO, and as poster and oral presentations at the 2017 AMCP Nexus Meeting; October 16-19, 2017; Dallas, TX. Part of this content was published as Hung's PhD dissertation.

REFERENCES

1. Mar J, Arrospide A, Comas M. Budget impact analysis of thrombolysis for stroke in Spain: a discrete event simulation model. *Value Health*. 2010;13(1):69-76.
2. Comas M, Arrospide A, Mar J, et al. Budget impact analysis of switching to digital mammography in a population-based breast cancer screening program: a discrete event simulation model. *PLoS One*. 2014;9(5):e97459.
3. Mehta D, Stolshek B, Agodoa I. Validation of a budget impact model for use of denosumab in postmenopausal women with osteoporosis [abstract]. *Value Health*. 2015;18(3):A157. Available at: [https://www.valueinhealthjournal.com/article/S1098-3015\(15\)00967-5/fulltext](https://www.valueinhealthjournal.com/article/S1098-3015(15)00967-5/fulltext). Accessed June 24, 2019.
4. Hung A, Vu Q. Validation of a budget impact model for a clinical management program for Zarxio [abstract U12]. *J Manag Care Spec Pharm*. 2017;23(3-a):S97. Available at: <https://www.jmcp.org/doi/pdf/10.18553/jmcp.2017.23.3-a.s1>. Accessed June 24, 2019.
5. Sullivan SD, Mauskopf JA, Augustovski F, et al. Budget impact analysis—principles of good practice: report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. *Value Health*. 2014;17(1):5-14.
6. Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices task force—7. *Value Health*. 2012;15(6):843-50.
7. Sargent RG. Verification and validation of simulation models. In: Johansson B, Jain S, Montoya-Torres J, Hagan J, Yücesan E, eds. *Proceedings of the 2010 Winter Simulation Conference*. Piscataway, NJ: IEEE; 2010:166-83. Available at: <https://www.informs-sim.org/wsc10papers/016.pdf>. Accessed June 25, 2019.
8. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics*. 2006;24(4):355-71.
9. Caro JJ, Briggs AH, Siebert U, Kuntz KM; ISPOR-SMDM Modeling Good Research Practices Task Force. Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force—1. *Value Health*. 2012;15(6):796-803.
10. Caro JJ, Eddy DM, Kan H, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health*. 2014;17(2):174-82.

11. Vemer P, Corro Ramos I, van Voorn GA, Al MJ, Feenstra TL. AdViSHE: a validation-assessment tool of health-economic models for decision makers and model users. *Pharmacoeconomics*. 2016;34(4):349-61.
12. Mauskopf JA, Sullivan SD, Annemans L, et al. Principles of good practice for budget impact analysis: report of the ISPOR task force on good research practices—budget impact analysis. *Value Health*. 2007;10(5):336-47.
13. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR task force on good research practices—modeling studies. *Value Health*. 2003;6(1):9-17.
14. Fildes R, Kourentzes N. Validation and forecasting accuracy in models of climate change. *Int J Forecast*. 2011;27(4):968-95.
15. Makridakis S, Taleb N. Living in a world of low levels of predictability. *Int J Forecast*. 2009;25(4):840-44.
16. Palmer AJ, Roze S, Valentine WJ, et al. Validation of the CORE Diabetes Model against epidemiological and clinical studies. *Curr Med Res Opin*. 2004;20(Suppl 1):S27-40.
17. Hoerger TJ, Segel JE, Zhang P, Sorensen SW. Validation of the CDC-RTI diabetes cost-effectiveness model. Methods report. RTI Press publication no. MR-0013-0909. September 2009. Available at: <https://www.rti.org/sites/default/files/resources/rti-publication-file-c23097bc-7c56-4baa-8221-6d57810d90ec.pdf>. Accessed June 18, 2019.
18. Yang J, Ning C, Deb C, et al. K-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy Build*. 2017;146(Suppl C):27-37.
19. Voyant C, Soubdhan T, Lauret P, David M, Muselli M. Statistical parameters as a means to a priori assess the accuracy of solar forecasting models. *Energy*. 2015;90(Part 1):671-79.
20. Wang Z. Stock returns and the short-run predictability of health expenditure: some empirical evidence. *Int J Forecast*. 2009;25(3):587-601.
21. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecast*. 2006;22(4):679-88.
22. Forecasting FAQs. In: Gilliland M, ed. *The Business Forecasting Deal*. Hoboken, NJ: John Wiley and Sons; 2012:193-246.
23. Hung A, Mullins CD, Slejko JF, Haines ST, Shaya F, Lugo A. Using a budget impact model framework to evaluate antidiabetic formulary changes and utilization management tools. *J Manag Care Spec Pharm*. 2019;25(3):342-49. Available at: <https://www.jmcp.org/doi/10.18553/jmcp.2019.25.3.342>.
24. Box GEP, Jenkins GM, Reinsel GC. *Time Series Analysis: Forecasting and Control*, 4th edition. Hoboken, NJ: John Wiley and Sons; 2008.
25. Choi Y, Navarro RP. Assessment of the level of satisfaction and unmet data needs for specialty drug formulary decisions in the United States. *J Manag Care Spec Pharm*. 2016;22(4):368-75. Available at: <https://www.jmcp.org/doi/10.18553/jmcp.2016.22.4.368>.
26. van de Vooren K, Duranti S, Curto A, Garattini L. A critical systematic review of budget impact analyses on drugs in the EU countries. *Appl Health Econ Health Policy*. 2014;12(1):33-40.
27. Orlewska E, Gulácsi L. Budget-impact analyses: a critical review of published studies. *Pharmacoeconomics*. 2009;27(10):807-27.
28. Neumann PJ. Budget impact analyses get some respect. *Value Health*. 2007;10(5):324-25.
29. Law AM. How to build valid and credible simulation models. In: Rossetti MD, Hill RR, Johansson B, Dunkin A, Ingalls RG, eds. *Proceedings of the 2009 Winter Simulation Conference*. 2009: Piscataway, NJ: IEEE; 2009:24-33. Available at: <https://www.informs-sim.org/wsc09papers/003.pdf>. Accessed June 25, 2019.

APPENDIX A Budget Impact Model Face Validity Questionnaire

Part 1. Selection of Experts

1. Who are the experts you have chosen?
2. Justify why they are considered experts.
3. Do the experts have any potential conflicts of interest with the model?

Part 2. Questions to Ask Experts

Structure/Conceptual Model

1. Does the model contain all aspects (e.g., interventions, medical and pharmacy resource utilization, health care system features) that are relevant to the decision according to the specified perspective?
2. Are these aspects represented and linked according to your best understanding?
3. Is the time horizon appropriate for the model's objective and long enough to account for all relevant aspects?

Data

1. Have the best available data sources been used (especially for current utilization, shifts in market share, and impact on other disease-related costs)? If data were not used for current utilization, shifts in market share, and/or impact on other disease-related costs, do predictions seem reasonable?
2. Is there potential bias in these data?
3. Are the data generalizable to the target population?

Results

1. Do the results match your expectations? If not, are they plausible?
 2. Repeat all questions above for each scenario examined in the budget impact model, including those in sensitivity analyses and subgroup analyses.
 3. Do sensitivity analyses include the most important scenarios?
-

APPENDIX B SGLT2I and GLP-1RA Formulary and Utilization Management Policy Changes²³

^aIncludes SU, TZD, DPP-4I, AGI, and GLN.

AGI=alpha-glucosidase inhibitor; DPP-4I=dipeptidyl peptidase-4 inhibitor; GLN=meglitinide; GLP-1RA=glucagon-like peptide-1 receptor agonist; MET=metformin; SGLT2I=sodium-glucose cotransporter-2 inhibitor; SU=sulfonylurea; TZD=thiazolidinedione; UM=utilization management.