Journal of Medical Economics

ISSN: 1369-6998 (Print) 1941-837X (Online) Journal homepage: informahealthcare.com/journals/ijme20

Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM)

Michael Willis, Christian Asseburg & Jianming He

To cite this article: Michael Willis, Christian Asseburg & Jianming He (2013) Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM), Journal of Medical Economics, 16:8, 1007-1021, DOI: 10.3111/13696998.2013.809352

To link to this article: https://doi.org/10.3111/13696998.2013.809352

Journal of Medical Economics Vol. 16, No. 8, 2013, 1007–1021

1369-6998
doi:10.3111/13696998.2013.809352

Article 0034.R1/809352

Original article

Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM)

Michael Willis
The Swedish Institute for Health Economics, Lund, Sweden

Christian Asseburg
ESiOR Oy, Kuopio, Finland

Jianming He
Janssen Global Services, LLC, Raritan, NJ, USA

Address for correspondence:
Michael Willis, The Swedish Institute for Health Economics, Box 2127, 220 02 Lund, Sweden. Tel: +46-46-329112; Fax: +46-46-122604; mw@ihe.se

Keywords:
Diabetes – Economic model – Validation

Accepted: 24 May 2013; published online: 26 June 2013
Citation: J Med Econ 2013; 16:1007–21

Abbreviations:
ACCORD, The Action to Control Cardiovascular Risk in Diabetes; ACE, angiotensin-converting-enzyme; ADVANCE, Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation; AE, adverse event; AP, angina pectoris; ASPEN, Atorvastatin Study for Prevention of coronary heart disease Endpoints in Non-insulin-dependent diabetes mellitus; BDR, background diabetic retinopathy; CARDS, Collaborative AtoRvastatin Diabetes Study; CHF, congestive heart failure; CVD, cardiovascular disease; DBP, diastolic blood pressure; ESRD, end-stage renal disease; GPR, gross proteinuria; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; ICER, incremental cost-effectiveness ratio; LDL, low-density lipoprotein; LEA, lower extremity amputation; LY, life year; MA, microalbuminuria; ME, macular edema; MI, myocardial infarction; MICRO-HOPE, MIcroalbuminuria, Cardiovascular, and Renal Outcomes. Heart Outcomes Prevention Evaluation; NMB, net monetary benefit; PDR, proliferative diabetic retinopathy; QALY, quality-adjusted life year; SBP, systolic blood pressure; SMDM, Society for Medical Decision Making; UKPDS, United Kingdom Prospective Diabetes Study; WESDR, Wisconsin Epidemiologic Study of Diabetic Retinopathy.

Abstract

Objective:
This study constructed the Economic and Health Outcomes Model for type 2 diabetes mellitus (ECHO-T2DM), a long-term stochastic microsimulation model, to predict the costs and health outcomes in patients with T2DM. Naturally, the usefulness of the model depends upon its predictive accuracy. The objective of this work is to present results of a formal validation exercise of ECHO-T2DM.

Methods:
The validity of ECHO-T2DM was assessed using criteria recommended by the International Society for Pharmacoeconomics and Outcomes Research/Society for Medical Decision Making (ISPOR/SMDM). Specifically, the results of a number of clinical trials were predicted and compared with observed study end-points using a scatterplot and regression approach. An *F*-test of the best-fitting regression was added to assess whether it differs statistically from the identity (45°) line defining perfect predictions. In addition to testing the full model using all of the validation study data, tests were also performed of microvascular, macrovascular, and survival outcomes separately. The validation tests were also performed separately by type of data (used vs not used to construct the model, economic simulations, and treatment effects).

Results:
The intercept and slope coefficients of the best-fitting regression line between the predicted outcomes and corresponding trial end-points in the main analysis were −0.0011 and 1.067, respectively, and the R^2 was 0.95. A formal *F*-test of no difference between the fitted line and the identity line could not be rejected ($p = 0.16$). The high R^2 confirms that the data points are closely (and linearly) associated with the fitted regression line. Additional analyses identified that disagreement was highest for macrovascular end-points, for which the intercept and slope coefficients were 0.0095 and 1.225, respectively. The R^2 was 0.95 and the estimated intercept and slope coefficients were 0.017 and 1.048, respectively, for mortality, and the *F*-test was narrowly rejected ($p = 0.04$). The sub-set of microvascular end-points showed some tendency to over-predict (the slope coefficient was 1.095), although concordance between predictions and observed values could not be rejected ($p = 0.16$).

Limitations:
Important study limitations include: (1) data availability limited one to tests based on end-of-study outcomes rather than time-varying outcomes during the studies analyzed; (2) complex inclusion and exclusion criteria in two studies were difficult to replicate; (3) some of the studies were older and reflect outdated treatment patterns; and (4) the authors were unable to identify published data on resource use and costs of T2DM suitable for testing the validity of the economic calculations.

Conclusions:
Using conventional methods, ECHO-T2DM simulated the treatment, progression, and patient outcomes observed in important clinical trials with an accuracy consistent with other well-accepted models. Macrovascular outcomes were over-predicted, which is common in health-economic models of diabetes (and may be related to a general over-prediction of event rates in the United Kingdom Prospective Diabetes Study [UKPDS] Outcomes Model). Work is underway in ECHO-T2DM to incorporate new risk equations to improve model prediction.

Introduction

Efficient allocation of healthcare resources requires economic evaluation to guide decision-making about which interventions to use[1]. For chronic and progressive disease, such as type 2 diabetes mellitus (T2DM), economic evaluation including estimation of cost-effectiveness generally requires the use of mathematical modeling to extrapolate the available data (e.g., treatment effects from short-run trials) to long-run health and economic outcomes using known physiological relationships[2,3].

We constructed a long-term, second order stochastic micro-simulation model of the treatment of T2DM, known as the Economic and Health Outcomes Model for T2DM (ECHO-T2DM), to estimate the cost-effectiveness of alternative diabetes treatments[4]. ECHO-T2DM incorporates key structural features from a number of well-known T2DM models (e.g., the NIH Model[5], the Core Diabetes Model[6], and DiDACT[7]) including development and progression of key micro- and macro-vascular complications, and mortality. A distinguishing feature of ECHO-T2DM is its comprehensive treatment sub-model, including a broad range of treatment consequences (both initially and over time), a flexible long-term treatment sequence and switching algorithm using treatment targets that can vary by treatment line, and a broad set of adverse events (AEs) making it suitable for modeling a broad group of T2DM agents in detail. ECHO-T2DM is depicted in Figure 1 and a more thorough technical description (including definitions of the health states) is presented in the Appendix.

Naturally, the usefulness of a model depends upon its ability to predict accurately the actual health and economic outcomes of patients in a real-life treatment setting. Model validation is a set of methods for judging predictive accuracy (i.e., how well a model agrees with observed outcomes in clinical practice)[8]. The latest joint International Society for Pharmacoeconomics and Outcomes Research/Society for Medical Decision Making (ISPOR/SMDM) good research guidelines describe five principal forms of validation. Face validity is the extent to which a model, its assumptions, and the applications for which it is used accurately reflect current scientific evidence (as judged by experts). Verification examines the extent to which the model calculations are correctly implemented. Cross-validation (often called convergent validity) consists of simulating identical scenarios with different models and comparing the results and examining differences. Simulating scenarios based on actual events that have occurred and assessing concordance is termed external validation, and called 'dependent' if the source data were used in model construction and 'independent' otherwise. Finally, predictive validation consists of external validation in which the study has not yet been conducted, ensuring that the external validation is completely independent.

Objective

The objective of this paper is to present results of a formal validation exercise of ECHO-T2DM.

Methods

We followed the ISPOR/SMDM principles of good practice[8]. The face validity of ECHO-T2DM has been evaluated throughout the model development process and upon completion in several ways. We obtained clinical expert feedback during the design and programming phase of the key model features, we have presented the model at numerous conferences, and we have participated with ECHO-T2DM at the Fifth and Sixth Mount Hood Challenges[9–11]. Predicting in advance the outcomes of a trial not yet conducted (i.e., 'predictive validation') has not been attempted.

Verification

The model has been thoroughly tested and de-bugged, including artificial simulations designed to reveal errors in both logic and programming (i.e., so-called 'stress tests'). Idiosyncratic results were investigated and any identified errors in programming or logic were corrected.

Cross-validation

We assessed cross-validity by replicating the intensive vs conventional blood glucose control analysis conducted and published by the NIH using the seminal model of T2DM[5,12] and then examining the degree of concordance between the two sets of predictions. Specifically, we loaded the model with the same distributions of baseline patient characteristics, applied identical treatment effects and assumptions about drift (annual evolution in subsequent years) in biomarker values over time, simulated for the same lifetime time horizon, and then extracted the same set of predicted cumulative incidence values. We compared them with the statistical approach outlined below. In addition, ECHO-T2DM has been subjected to cross-validation as part of the Fifth and Sixth Mount Hood Challenges, in which a large number of modeling groups predicted end-points for a number of standardized scenarios, and those results will be disseminated separately[13].

External validation (dependent and independent)

Like previous work[14,15], we assessed the external validity of ECHO-T2DM by simulating the key features of a broad set of clinical trials and then compared the model predictions with the corresponding actual observed outcomes for a variety of clinical end-points. The choice of studies is

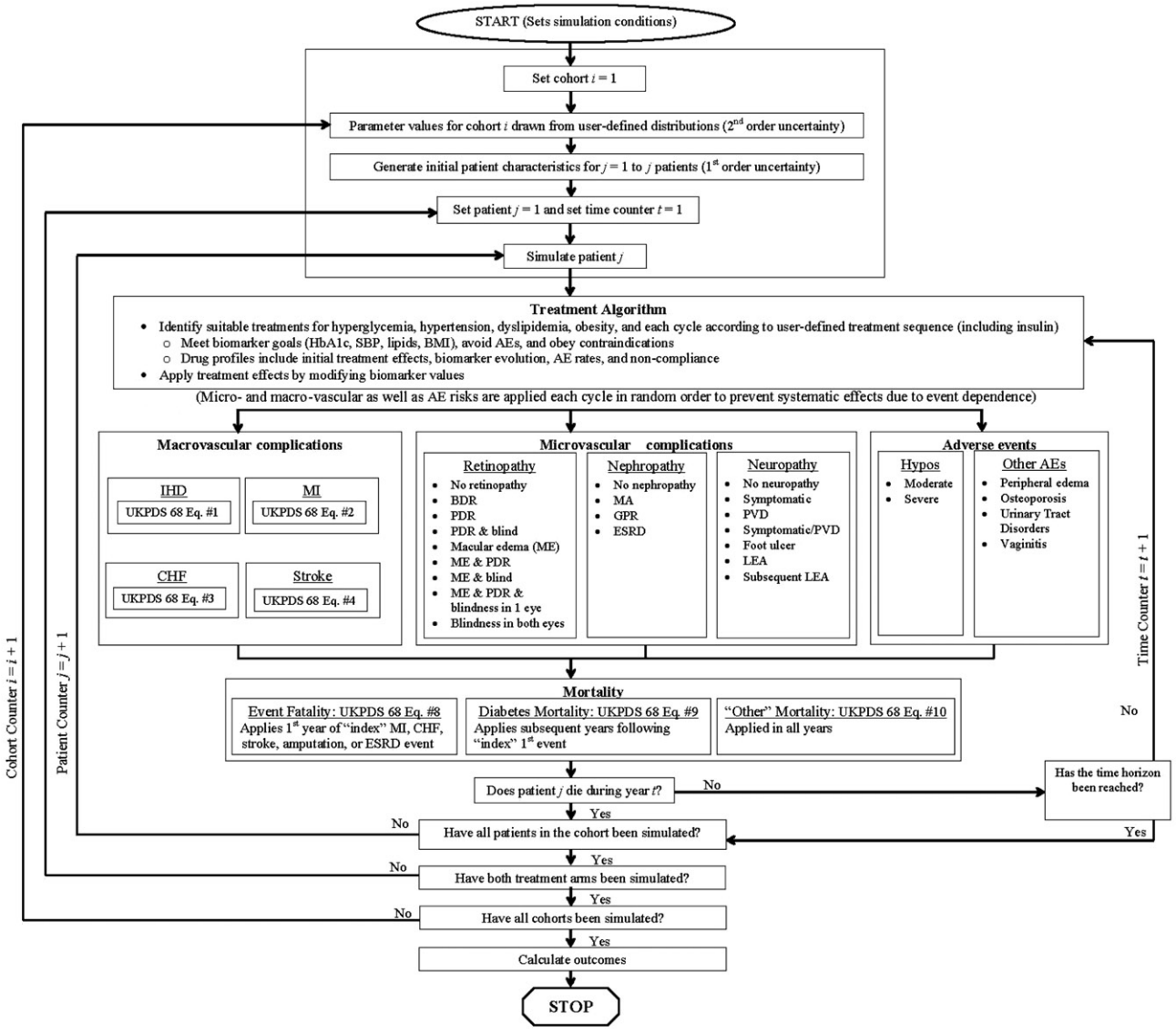

Figure 1. ECHO-T2DM schematic. SBP, systolic blood pressure; BMI, body mass index; AE, adverse event; IHD, ischemic heart disease; UKPDS, UK prospective diabetes study; MI, myocardial infarction; CHF, congestive heart failure; BDR, background diabetic retinopathy; PDR, proliferative diabetic retinopathy; ME, macular edema; MA, microalbuminuria; GPR, gross proteinuria; ESRD, end-stage renal disease; PVD, peripheral vascular disease; LEA, lower extremity amputation.

important and, for a multi-application model, should include data from differing settings. Published studies were chosen for validation based on importance, relevance, and replicability (i.e., the publicly available data on baseline patient characteristics and outcomes were sufficiently comprehensive to allow modeling). With inspiration from Eddy and Schlessinger[15], Palmer *et al.*[14], and The Mount Hood 4 Modeling Group[9], five published studies were selected in addition to the NIH cross-validation study mentioned above (see Table 1).

Two of these studies were used in construction of the model and form a sub-set of dependent external validation studies; the UK Prospective Diabetes Study (UKPDS) is the source of macrovascular and mortality risks[16,17], and the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) is the source of many of the microvascular transition probabilities[18]. The other three studies were not part of model construction, and formed the independent external validation sub-set. CARDS[19,20], which investigated the role of lipid-lowering therapy in preventing CVD in a cohort of T2DM patients, was recommended by The Mount Hood 4 Modeling Group; MICRO-HOPE[21,22], a large interventional trial of ACE Inhibitors in T2DM patients with high risk of CVD, was used in validation of ARCHIMEDES[15]; and the Osaka study of T2DM survival, an observational 15-year study of nearly 2000 patients in Japan[23], was used in validation of the CORE model[14]. ECHO-T2DM has also been validated

Table 1. Validation simulation analyses.

Trial	Population	Duration (years)	Treatment groups	Sample size
UKPDS [17]	Interventional study of newly diagnosed T2DM, aged 25–65, no MI in previous year and no current AP or CHF in the UK.	12	Conventional/Intensive	3867
CARDS [20]	Interventional study of relatively healthy T2DM.	3.9	Placebo/Atorvastatin	2838
MICRO-HOPE [22]	Interventional study of T2DM with high risk of CVD.	4	Placebo/Ramipril	3657
WESDR [18]	Observational study of DM diagnosed at or after 30 years of age in an 11-county area in southern Wisconsin.	Up to 30	Older Onset Group, Older Onset With Insulin, Older Onset W/out Insulin, Mixed Newly Diagnosed	1370, 143, 296, 639
EASTMAN [12]	Seminal model of T2DM. Comparison of conventional with intensive care (goal of normoglycemia) for newly diagnosed T2DM.	Lifetime	Standard/Comprehensive	Model
Sasaki (OSAKA) [23]	Long-term follow-up mortality study in Osaka, Japan. Survival in age groups: 35–44, 45–54, 55–64, and 65+.	Up to 20	Observational	1939

T2DM, type 2 diabetes mellitus; MI, myocardial infarction; AP, angina pectoris; CHF, congestive heart failure; CVD, cardiovascular disease.

against the recent ADVANCE, ACCORD, and ASPEN trials as part of the fifth Mount Hood Challenge; the results and those of the other T2DM models are published separately and are not reported here[13].

Statistical methods

A conceptual summary of methods can be found in Figure 2. We loaded the ECHO-T2DM model with the same distributions of baseline patient characteristics, reflecting study inclusion and exclusion criteria. Complicated inclusion and exclusion criteria in two of the clinical trials, CARDS[20] and MICRO-HOPE[22], could only be approximated because of mismatches with health states. (Individuals were included in CARDS who had T2DM diagnosed at least 6 months before study entry and at least one of the following: history of hypertension [defined as receiving antihypertensive treatment or SBP ≥140 mmHg or DBP ≥90 mmHg on at least two successive occasions], retinopathy [any retinopathy, maculopathy, or previous photocoagulation], microalbuminuria or macroalbuminuria [positive Micral or other strip test, albumin creatinine ratio ≥2.5 mg/mmol, or albumin excretion rate on time collection of ≥20 μg/min, all on ≥2 successive occasions], or currently smoking. Exclusion criteria included past history of myocardial infarction; angina; coronary vascular surgery; cerebrovascular accident; or severe peripheral vascular disease [warranting surgery]; mean serum LDL-cholesterol and triglyceride concentrations during baseline visits ≥4.14 mmol/L and 6.78 mmol/L, respectively; plasma creatinine concentration ≥150 μmol/L; HbA1c ≥12%; and <80% compliance with placebo in the baseline phase. Individuals were included in MICRO-HOPE who had diabetes, were aged 55 years or older, and who had a history of cardiovascular disease [coronary artery disease, stroke, or peripheral vascular disease] or at least one other cardiovascular risk factor [total cholesterol >5.2 mmol/L, HDL cholesterol ≤0.9 mmol/L, hypertension, known microalbuminuria, or current smoking]. Key exclusion criteria were dipstick-positive proteinuria or established diabetic nephropathy, other severe renal disease, hyperkalemia, congestive heart failure, low ejection fraction [<0.4], uncontrolled hypertension, recent myocardial infarction or stroke [<4 weeks], and use or hypersensitivity to vitamin E or ACE inhibitors.) For parameters where no data were publicly available, we used corresponding data from a suitable alternative (often the UKPDS because it is well-known, the data are of high quality, and it is well-documented so even values for minor parameters could often be identified; for example, HbA1c evolution was not obtainable for each of the validation studies. In line with convention, the annual upward drift observed in the UKPDS of 0.15% per year was then assumed). The study treatment protocol and resulting treatment effects (one time change and subsequent annual drift) and AE rates were then applied and the model was simulated for a time horizon equal to the length of the trial.

We extracted cumulative incidences for the following end-points: stroke; myocardial infarction (MI); angina pectoris (AP); congestive heart failure (CHF); a composite end-point consisting of stroke, MI, and AP/cardiac arrest; microalbuminuria (MA); gross proteinuria (GPR); end-stage renal disease (ESRD); background and proliferative retinopathy; blindness; neuropathy; lower extremity amputation (LEA); and survival/mortality (To match the

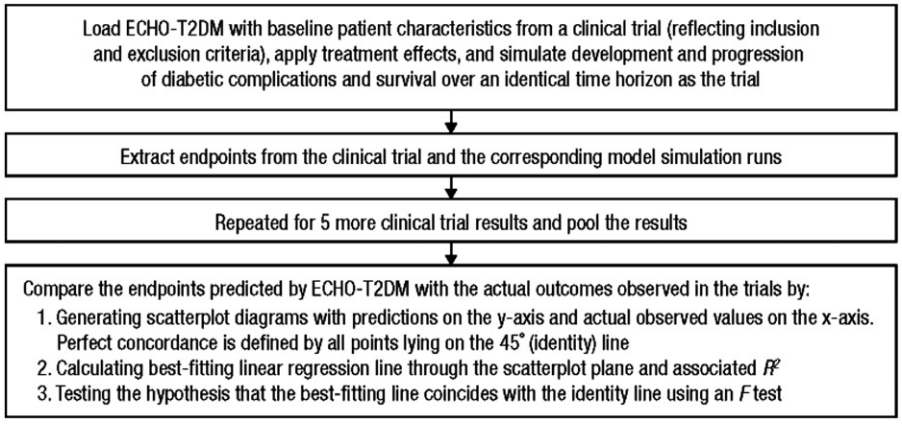

Figure 2. Conceptual summary of methods.

definitions used in the clinical trials, we were required in some cases to convert model results *ex post*. For example, our model does not generate the composite primary cardiovascular end-point in the CARDS trial, so we computed it separately assuming independence of occurrence. In other cases, the study end-points were the cumulative incidences of *new* events only [e.g., BDR and PDR in WESDR], thus adjustments were made to net out the effects of baseline disease patterns assuming event independence.). Not all end-points were publicly reported for each of the validation studies. In total, 80 validation end-points across the six studies (five external validation and one cross-validation) were included, reflecting different randomization arms, patient cohorts, and end-points.

A number of approaches have been used empirically to evaluate concordance between T2DM model predictions and actual observed outcomes. Eddy and Schlessinger[15] tested the validity of Archimedes Diabetes Model, a mathematical model that includes the physiology of diabetes (both Type 1 and Type 2) and its complications, against randomized controlled trials by comparing Kaplan-Meier curves with published results for each of the outcomes reported in the trials using a log-rank test of no statistically significant difference at the 5% level (separately for each end-point). In order to gain an overview of the full set of validation exercises, the predicted cumulative incidences were also plotted graphically against the results of the actual trials, and the correlation coefficient (r) was calculated. The identity (45°) line defines perfect agreement between the model estimates and the corresponding trial results. Palmer *et al.*[14] and Hoerger *et al.*[24] tested the validity of the CORE and the CDC-RTI Diabetes Cost-Effectiveness Model T2DM models, respectively, by plotting predicted outcomes vs actual observed outcomes from a set of clinical trials (some non-interventional) and the NIH analysis, estimating the best-fitting line emanating from the origin and calculating the corresponding R^2 to illustrate the degree to which the measures co-vary. While not as strong as the log-rank test employed by Eddy and Schlessinger[15], this approach does not require data for the entire time-path of the simulation (a graphical analysis was provided for some trials where data permitted). Mueller *et al.*[25] employed a third approach to quantify the comparison; they defined model validity as mean simulated event rates that fall within a tolerance of 10% of the mean trial event rates.

We assessed concordance of ECHO-T2DM predictions and the actual observed outcomes using the scatterplot and linear regression approach adopted by Palmer *et al.*[14]. Predicted results were plotted against actual results using a two-dimensional scatterplot and then linear regression (including an intercept) was used to estimate a best-fitting line using STATA (College Station, TX), defined as

$$P_i = \beta_0 + \beta_1 * O_i + \varepsilon_i,$$

where P_i is the predicted cumulative incidence for the ith end-point, O_i is the observed value in the actual study for the ith end-point, and β_0 and β_1 are the intercept and slope coefficients, respectively. ε is the disturbance term. The standard errors were estimated using the heteroskedasticity-consistent Huber-White estimator to account for correlation between end-points in the same clinical trial (e.g., above-average rates of MI may accompany above-average rates of stroke).

Because a high coefficient of determination, R^2, indicates only that the points in the scatterplot lie collectively close to the predicted line (i.e., that they are linear), we extend the above approach by testing formally for a statistically significant difference between the estimated coefficients and the identity line (which corresponds to perfect concordance between predicted and observed

values). Specifically, this consists of a joint F-test of the null hypothesis:

$$H_0 : \beta_0 = 0 \text{ and } \beta_1 = 1.$$

The null hypothesis that the model predictions and the actual trial outcomes agree was tested at the 5% level of statistical significance.

Rejection of the null hypothesis implies that the best-fitting line through the points defined by the model predictions and observed values is not the identity line. Non-rejection of the null hypothesis implies that the best-fitting line may be the identity line, but it does not imply that the points defined by the model predictions and observed values actually lie on (or even close to) the best-fitting line. The R^2 value measures the closeness of the points to the linear regression line; an R^2 equal to 1 (or close to 1) indicates that the validation data-points lie necessarily on (or close to) the fitted regression line. Failure to reject the F-test, together with a high R^2, thus, indicate concordance between model predictions and observed outcomes (that is, the data points lie close to the best fitting regression line, which cannot be statistically distinguished from the identity line).

Analyses

Our main analysis includes all validation end-points, including both the cross-validity and the external validity studies. In addition, we also performed validation tests on specific parts of the model separately (microvascular events, macrovascular events, and survival). To further examine the performance of the model, we performed validation tests separately for data that were used in constructing the model (dependent external), data that were not used in constructing the model (independent external), and data estimated by another economic model (cross-validity). Finally, we also repeated the analysis for treatment-related differences in outcomes to assess how well the health-economic model predicts the consequences of different treatment strategies. (Because studies without a comparator are excluded and because one treatment and one control end-point are required to compute each treatment effect, the sample size is smaller than for the other analyses.)

Results

The main results using the full validation data set are presented in Table 2. The actual and model predicted outcomes (cumulative incidences) are also presented graphically in Figure 3. Visually, the points lie close to the identity line. The intercept and slope coefficients of the best-fitting regression line are −0.001 and 1.067, respectively. The F-test fails to reject the null hypothesis of agreement between the model and trial outcomes ($p = 0.16$) and the R^2, 0.95, is high, confirming that the points lie collectively close to the regression line.

Additional analyses

A tendency to over-predict outcomes in the sub-set of macrovascular end-points ($n = 18$) can be seen in Figure 4. The slope coefficient is 1.225 and the null hypothesis of model and trial agreement can also be rejected ($p = 0.003$). The R^2 of 0.87 confirms the relative linearity of the relationship. The result was largely driven by end-points from two of the validation studies (CARDS and MICRO-HOPE), for which ECHO-T2DM over-predicted by ~50%.

In Figure 5, the regression line appears to fit the points well for the survival/mortality end-points, with a few outliers noticeable in the middle of the distribution ($n = 24$). The intercept and slope coefficients are 0.017 and 1.048, respectively, and the R^2 was 0.95. The null hypothesis of model agreement, however, was narrowly rejected ($p = 0.04$).

While ECHO-T2DM tended to over-estimate the rate of microvascular events as well ($n = 38$), with intercept and slope coefficients of −0.023 and 1.095, respectively, we fail to reject the null hypothesis of agreement between model predictions and the trial results ($p = 0.12$; Figure 6). The relatively high R^2 of 0.90, moreover, indicates that the data points lie collectively close to the best fitting regression line.

As expected, the model performed somewhat better for the sub-set of studies used in model construction than for studies not used in model construction (which, moreover, include two cardiovascular intervention studies that were difficult to simulate). The outcomes for the sub-set of dependent external validation end-points ($n = 32$) are presented graphically in Figure 7. The intercept and slope coefficients are −0.002 and 1.024, respectively, and the null hypothesis cannot be rejected ($p = 0.82$), indicating that the best-fitting regression line cannot be distinguished from the identity line. The R^2, 0.95, is high, indicating that the data points lie collectively close to the best fitting regression line (which is itself statistically indistinguishable from the identity line).

The outcomes for the sub-set of independent external validation end-points are presented graphically in Figure 8 ($n = 30$). The intercept and slope coefficients are 0.015 and 1.072, respectively, and the R^2 is 0.97 for the best-fitting regression line, and the null hypothesis of model and trial agreement can be rejected ($p = 0.001$). While most of the points lie relatively

Table 2. Actual and predicted study outcomes.

Trial	End-point	Years	Treatment group	Actual study cumulative incidence	Predicted cumulative incidence	Difference
UKPDS	Stroke	12	Conventional	0.061	0.101	−0.040
			Intensive	0.061	0.093	−0.032
	MI	12	Conventional	0.190	0.225	−0.035
			Intensive	0.160	0.212	−0.052
	AP	12	Conventional	0.067	0.093	−0.026
			Intensive	0.068	0.088	−0.020
	CHF	12	Conventional	0.033	0.081	−0.048
			Intensive	0.030	0.077	−0.047
	Microalbuminuria	12	Conventional	0.340	0.223	0.117
			Intensive	0.230	0.208	0.022
	Proteinuria	12	Conventional	0.103	0.083	0.020
			Intensive	0.068	0.066	0.002
	ESRD	12	Conventional	0.008	0.002	0.006
			Intensive	0.006	0.002	0.004
	Retinopathy	12	Conventional	0.490	0.369	0.121
			Intensive	0.390	0.270	0.120
	Survival	12	Conventional	0.850	0.828	0.022
			Intensive	0.850	0.839	0.011
	Blindness in one eye	12	Conventional	0.035	0.003	0.032
			Intensive	0.029	0.002	0.027
CARDS	Primary end-point (fatal and non-fatal stroke, MI, and AP + cardiac arrest)	4	Placebo	0.090	0.125	−0.035
			Atorvastatin	0.058	0.097	−0.039
	Stroke	4	Placebo	0.028	0.031	−0.003
			Atorvastatin	0.015	0.028	−0.013
	Death from any cause	4	Placebo	0.058	0.076	−0.018
			Atorvastatin	0.043	0.069	−0.026
MICRO-HOPE	MI	4	Placebo	0.129	0.172	−0.043
			Ramipril	0.102	0.166	−0.064
	Stroke	4	Placebo	0.061	0.055	0.006
			Ramipril	0.042	0.051	−0.009
	Mortality	4	Placebo	0.140	0.182	−0.042
			Ramipril	0.108	0.177	−0.069
	CHF	4	Placebo	0.045	0.026	0.019
			Ramipril	0.045	0.025	0.020
WESDR	Mortality	4	Older Onset Group	0.240	0.203	0.037
	Mortality	10	Older Onset Group	0.550	0.571	−0.021
	New incidence of BDR	4	Older Onset With Insulin	0.474	0.636	−0.162
		4	Older Onset Without Insulin	0.344	0.344	0.000
	New incidence of PDR	4	Older Onset With Insulin	0.074	0.031	0.043
			Older Onset Without Insulin	0.023	0.012	0.011
	BDR	10	Mixed Newly Diagnosed	0.600	0.717	−0.117
	BDR	20	Mixed Newly Diagnosed	0.720	0.797	−0.077
	BDR	30	Mixed Newly Diagnosed	0.740	0.808	−0.068
	PDR	4	Mixed Newly Diagnosed	0.020	0.008	0.012
	PDR	16	Mixed Newly Diagnosed	0.050	0.097	−0.047
	PDR	30	Mixed Newly Diagnosed	0.175	0.120	0.055
EASTMAN	BDR	Lifetime	Standard Care	0.790	0.892	−0.102
			Comprehensive Care	0.270	0.255	0.015
	ME	Lifetime	Standard Care	0.520	0.645	−0.125
			Comprehensive Care	0.150	0.165	−0.015
	PDR	Lifetime	Standard Care	0.190	0.070	0.120
			Comprehensive Care	0.010	0.008	0.002
	Blindness	Lifetime	Standard Care	0.190	0.056	0.134
			Comprehensive Care	0.050	0.007	0.043
	MA	Lifetime	Standard Care	0.530	0.557	−0.027
			Comprehensive Care	0.320	0.312	0.008
	GPR	Lifetime	Standard Care	0.400	0.229	0.171
			Comprehensive Care	0.050	0.014	0.036
	ESRD	Lifetime	Standard Care	0.170	0.076	0.094
			Comprehensive Care	0.020	0.005	0.015

(continued)

Table 2. Continued.

Trial	End-point	Years	Treatment group	Actual study cumulative incidence	Predicted cumulative incidence	Difference
	Neuropathy	Lifetime	Standard Care	0.310	0.461	−0.151
			Comprehensive Care	0.100	0.291	−0.191
	LEA	Lifetime	Standard Care	0.150	0.214	−0.064
			Comprehensive Care	0.050	0.140	−0.090
Sasaki (OSAKA)	Survival (ages 35–44)	4	Observational	0.970	0.988	−0.018
		10	Observational	0.900	0.957	−0.057
		14	Observational	0.800	0.910	−0.110
		20	Observational	0.700	0.734	−0.034
	Survival (ages 45–54)	4	Observational	0.950	0.976	−0.026
		10	Observational	0.810	0.909	−0.099
		14	Observational	0.690	0.809	−0.119
		20	Observational	0.580	0.500	0.080
	Survival (ages 55–64)	4	Observational	0.850	0.949	−0.099
		10	Observational	0.640	0.801	−0.161
		14	Observational	0.520	0.609	−0.089
		20	Observational	0.370	0.220	0.150
	Survival (ages 65+)	4	Observational	0.720	0.871	−0.151
		10	Observational	0.330	0.512	−0.182
		14	Observational	0.180	0.236	−0.056
		20	Observational	0.080	0.027	0.053

MI, myocardial infarction; AP, angina pectoris; CHF, congestive heart failure; ESRD, end-stage renal disease; BDR, background diabetic retinopathy; PDR, proliferative diabetic retinopathy; ME, macular edema; MA, microalbuminuria; GPR, gross proteinuria; LEA, lower extremity amputation.

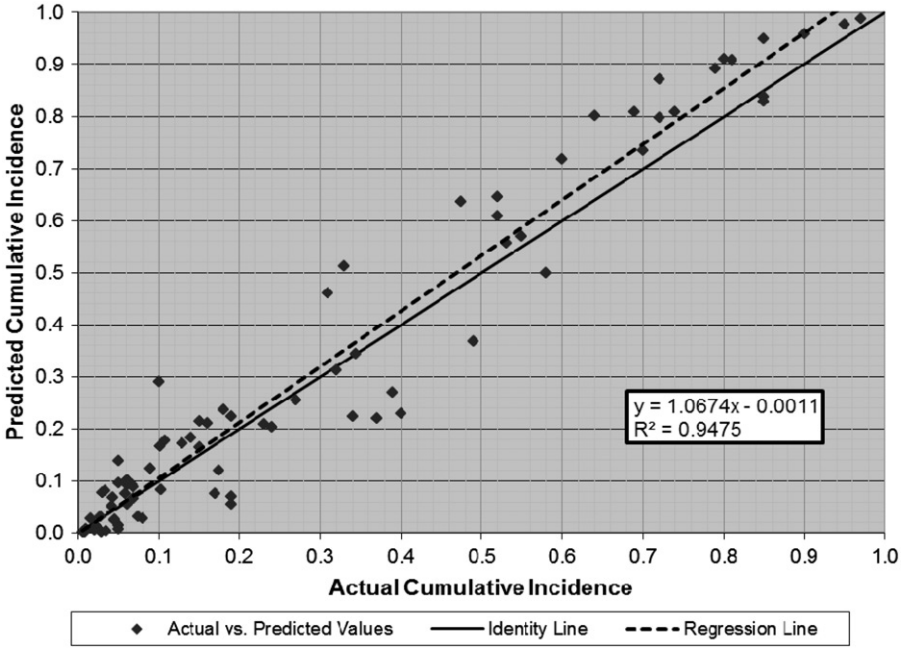

Figure 3. Predicted vs actual mean cumulative incidence (all outcomes, full validation data set).

close to the identity line, there is a tendency for the predicted values to exceed the actual trial values and several outliers are noticeable.

The results of the test of cross-validity between ECHO-T2DM and the NIH model of T2DM are presented in Figure 9 ($n = 18$). The intercept and slope coefficients are −0.014 and 1.090, respectively, and the null hypothesis of model agreement cannot be rejected ($p = 0.56$). The R^2 of 0.85 is lower than for the tests vs actual clinical studies, however, and many of the points lie noticeably distant from the best-fitting regression line.

The validation outcomes for the analysis of treatment effects from the sub-set of comparator studies

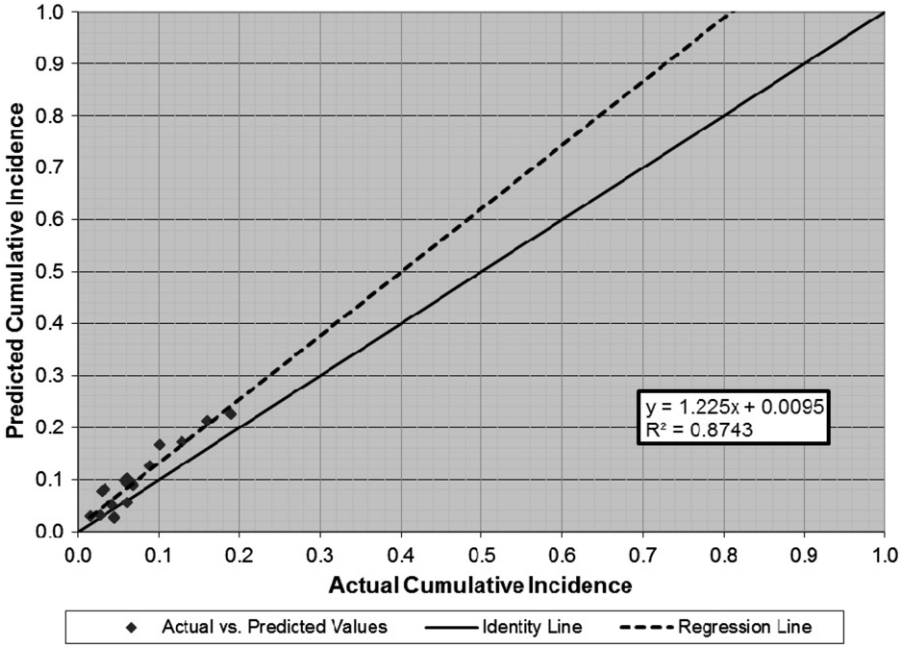

Figure 4. Predicted vs actual cumulative incidence (macrovascular end-points, full validation data set).

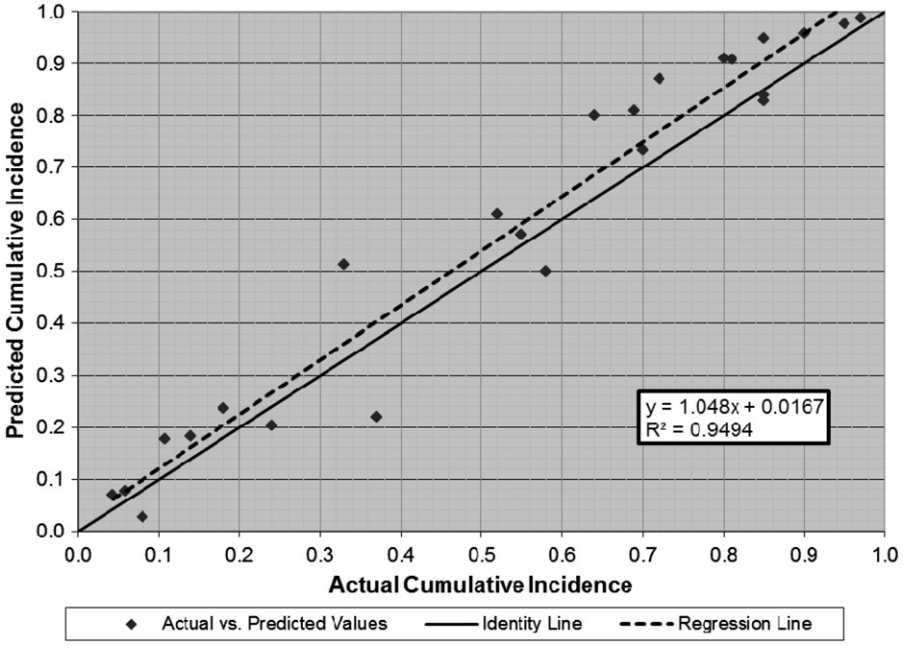

Figure 5. Predicted vs actual cumulative incidence (survival end-points, full validation data set).

are presented graphically in Figure 10. This analysis is limited to the 26 end-points in the UKPDS, CARDS, MICRO-HOPE, and the NIH Model. Despite fewer observations and the greater challenge of predicting treatment effects, the results look similar to many of the other analyses. The intercept and slope coefficients are −0.025 and 1.079, respectively, although the null hypothesis of agreement between the model predictions and the trial results is narrowly rejected ($p = 0.04$).

Discussion

Principles of good practice in health economic modeling emphasize the importance of extensive validity

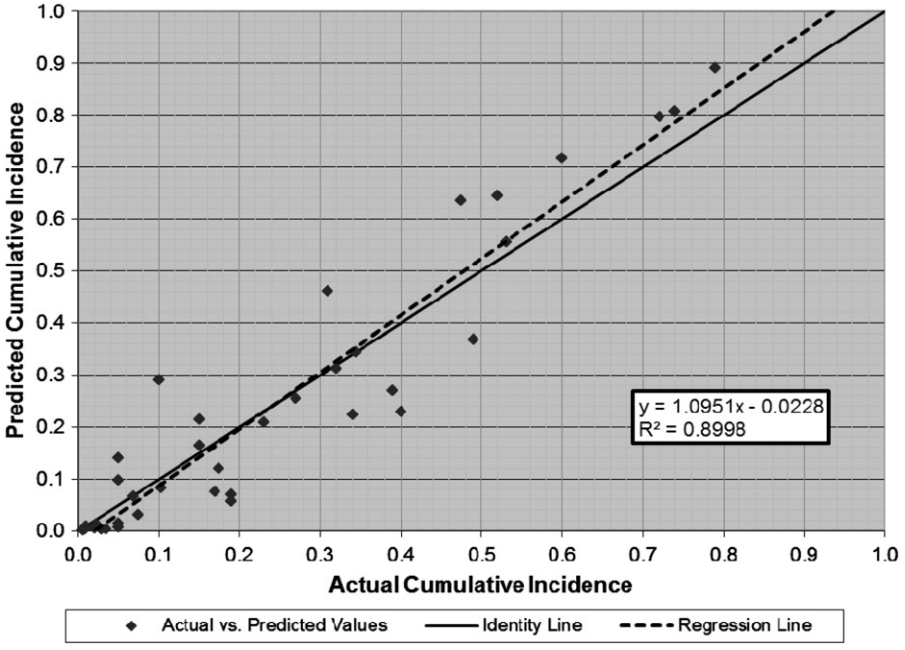

Figure 6. Predicted vs actual cumulative incidence (microvascular end-points, full validation data set).

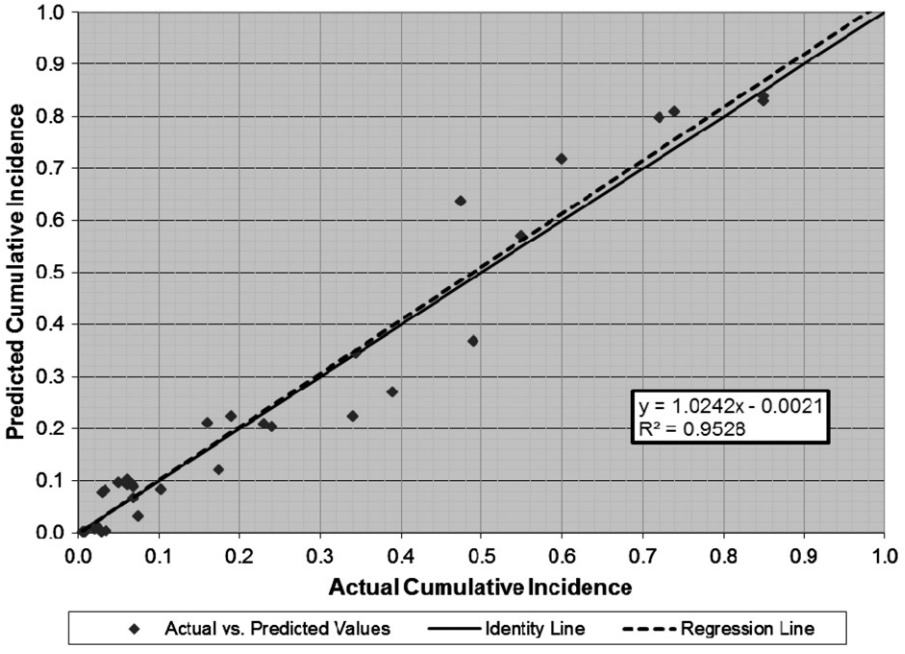

Figure 7. Predicted vs actual cumulative incidence (all outcomes, dependent external validation data set).

testing vs data from actual clinical trials, where close reproduction of a broad set of results is viewed as evidence of model validity[8]. Ideally models should be validated against clinical studies not used in model development (independent external validation) in order to provide a good indication of the accuracy of the model[26]. Moreover, if the modelers wish to claim a model as a general 'diabetes model' (i.e., suitable for more than just simulation of a particular clinical study, i.e. a 'multi-application' model), the model should not be calibrated to fit the validation exercises individually[26].

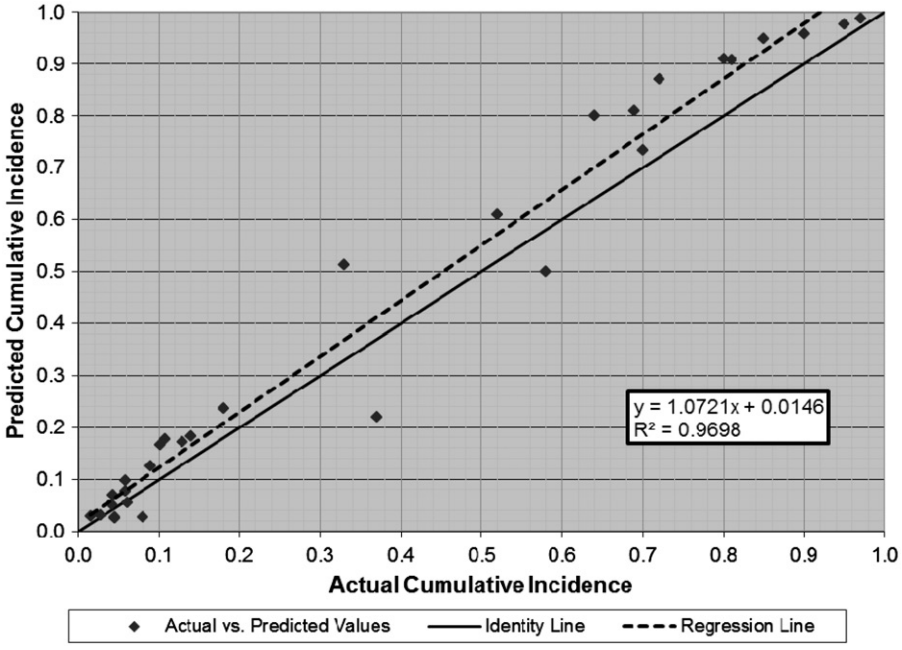

Figure 8. Predicted vs actual cumulative incidence (all outcomes, independent external validation data set).

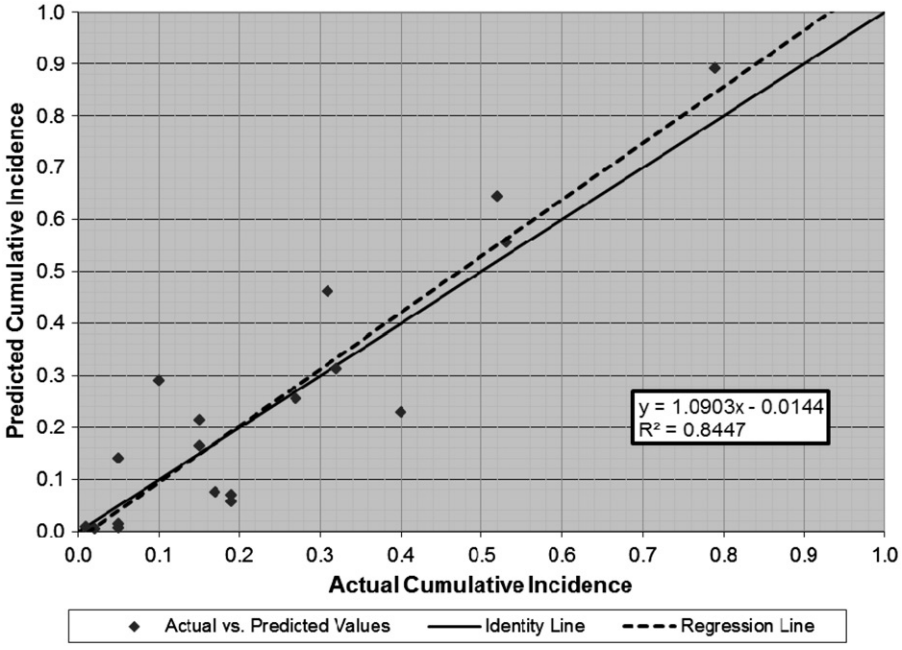

Figure 9. Predicted vs actual cumulative incidence (all outcomes, cross-validity endpoints).

We subjected ECHO-T2DM to extensive validation testing. Pairs of predicted outcomes and actual trial endpoints lie generally close to the identity line and the best fitting regression has intercept and slope coefficients close to 0 and 1, respectively, and a high R^2 (0.95). Unlike previous validation examples with economic models of T2DM, we also subjected the predictions to a formal test of whether the intercept and slope coefficients were statistically different from the identity line, finding no evidence of a difference in the main analysis. Taken together, the high R^2 and failure of the F-test to reject the null hypothesis indicate that the data points lie collectively close to the best fitting regression line and that the regression line cannot be distinguished statistically

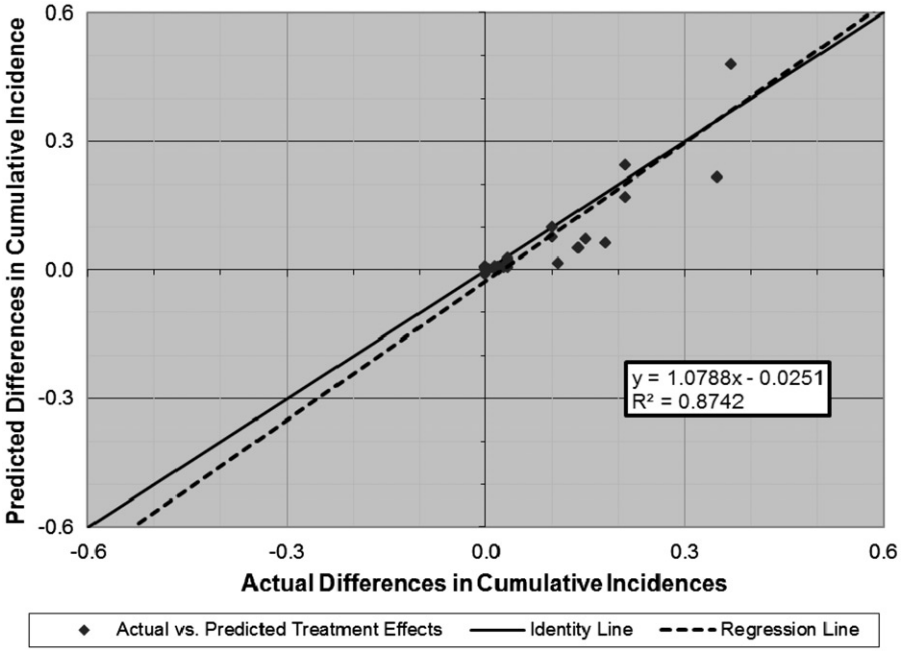

Figure 10. Predicted vs actual differences in cumulative incidence (all outcomes, comparative studies).

from the identity line and suggest that ECHO-T2DM predicted a wide range of key clinical outcomes for T2DM with reasonable accuracy.

Analysis of results by type of outcome suggests that the model performs quite well for microvascular complications and reasonably well for survival. There was a small tendency to over-predict complications, much of which can be traced to macrovascular outcomes and in particular to two studies that were difficult to mimic (CARDS and MICRO-HOPE). In particular, ECHO-T2DM over-estimated the event rates in the sub-set of end-points from CARDS and MICRO-HOPE by more than 50% (the R^2 was 0.88). The best-fitting regression line for the remaining macrovascular end-points (exclusively based on UKPDS data), in contrast, had slope and intercept coefficients of 0.037 and 1.011, respectively, and an R^2 of 0.97. This tendency to over-estimate macrovascular event rates has been seen elsewhere. Palmer and The Mount Hood Modeling Group[13], for example, over-estimated each of the four macrovascular end-points in their study (by an average of 18%) and, with one exception, each of the models participating in the Mt. Hood Challenge 4[9] over-estimated the rate of acute coronary events from the CARDS study. Interestingly, all participants overestimated the rate of strokes in the intervention arm but under-estimated them in the control arm.

The over-estimation of the macrovascular event rates may have several causes. First, the relatively serious macrovascular cumulative incidence rates were confined to a small range (0–0.2) of the interval between 0 and 1, which limits the precision with which the regression coefficients can be estimated (i.e., greater span reduces the standard error). In contrast, the sample points in the survival and microvascular complications analyses spanned nearly the full range from 0 to 1. Second, we use the UKPDS risk engine[16] to estimate macrovascular event rates, which has been shown to over-estimate event rates in actual patients[27–29]. Third, the sub-set of macrovascular end-points was influenced heavily by two studies for which it was particularly difficult to define the initial patient characteristics. Both CARDS[19] and MICRO-HOPE[22] had complicated, multi-dimensional inclusion and exclusion criteria, thus hampering our ability to generate a representative hypothetical cohort. Indeed, while the sample size is too small to draw reliable inferences, ECHO-T2DM over-predicted the macrovascular outcomes for CARDS and MICRO-HOPE by ~50%.

As expected, the model performed somewhat better for the sub-set of studies used in model construction (i.e., the 'dependent external validation'), where the best fitting regression line was not statistically different from the identity line, than for studies not used in model construction (i.e., the 'external validation'). Nevertheless, despite the inclusion of CARDS and MICRO-HOPE in the dependent external validation data set, the results were clearly respectable.

ECHO-T2DM also did a reasonable job in predicting treatment effects (i.e., the difference in event rates by treatment arm), with most points lying near the identity line and an R^2 of 0.874. There were several outliers in which ECHO-T2DM under-predicted the treatment effect, however, and the null hypothesis that the

predictions agree with the actual results could be rejected ($p = 0.04$). These outliers were largely related to the comparison vs the NIH model, and it is not clear which, if either, of the models generated correct estimates.

While the underlying analysis data sets varied, complicating direct comparison of these results to the validation testing of other models of T2DM, the results reported here accord reasonably well with the results observed elsewhere. Palmer *et al.*[14] tested the validity of the Core Diabetes Model and estimated a linear slope coefficient of 1.02 and an R^2 of 0.92 for an analysis combining both T2DM and Type 1 DM outcomes. The R^2 was slightly lower (0.89) for T2DM, but the corresponding slope coefficient was not reported. The estimated slope coefficient was even closer to 1 for the CDC model[24], 1.001 with an R^2 of 0.992 for dependent external validation studies and 0.991 with an R^2 of 0.969 for independent external validation studies. Zhou *et al.*[30] tested the external validity of the Michigan Model using the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) T2DM cohort, a population-based study in southern Wisconsin conducted largely in the 1980s. It is unclear whether WESDR data were used in model construction. While the authors conclude that 'the model is internally valid' (p. 2861), they do not present empirical metrics by which to assess validity. In a follow-up validation analysis of a revised version of the Michigan Model[31], the authors present internal validation results vs five end-points in the UKPDS[17] using relative errors. A summary measure was not calculated. Mueller *et al.*[25] tested the validity of the EAGLE model against outcomes in the key studies used in constructing the model (i.e., internal validation only) using differences in the predicted and actual results to re-calibrate the risk equations. Model validity was defined as mean simulated event rates that fall within a tolerance of 10% of the mean trial event rates. By definition, model predictions fell within ±10% for all of the outcomes studied for T2DM. For four of the 17 outcomes, however, the intervals formed by the ratios of the iteration with the lowest value to the mean value and the iteration with the highest to the mean value, which form a sort of quasi-confidence-interval, did not contain the value of the corresponding study outcome. Using the arguably strongest statistical test, Eddy and Schlessinger[15] found concordance (defined as no statistically significant difference at the 5% significance level) between model-predicted Kaplan-Meier curves and the time path of trial outcomes for 71 of 74 exercises. The correlation coefficient, r, was 0.99.

Strengths

The methodological approach presented here is straightforward and generally suitable for validation against any type of end-point (including treatment effects), and the data requirements are limited to the summary results commonly published from clinical trials. Moreover, the approach extends methods others have used[14,24], allowing comparison with earlier results but including additionally a formal hypothesis test to improve our ability to assess whether our results differ statistically from the clinical trial results.

Empirically, the study benefited from a diverse set of end-points spanning a number of different micro- and macro-vascular events, plus survival. These end-points, moreover, came from studies with different objectives, patient recruitment (including country setting and disease duration at baseline), intervention and randomization, and follow-up lengths. Several of the studies (e.g., CARDS and MICRO-HOPE) were particularly difficult to replicate, creating an additional source of uncertainty and, thus, biasing the tests against finding agreement.

Weaknesses

While the approach can be applied without access to detailed study results (which are often not publically available) necessary for the more powerful log-rank test (see Eddy and Schlessinger[15]), it generates two summary measures of model validity that must be considered together (p-value from the formal hypothesis test and R^2 of the linear regression), which inhibits unambiguous interpretation. A high R^2 implies only that the data points lie collectively closely to the best-fitting regression line, not that this regression line coincides with the identity line. Rejection of the F-test implies only that the best-fitting regression line cannot be distinguished statistically from the identity line, but the data points are not required to lie proximate to the regression line. Additionally, rejection of the F-test may simply be due to large sample size; ideally, power calculations should be used to control the risk of Type 1 error (i.e., rejecting an accurate model).

The approaches reviewed and presented here consider only mean values and, thus, ignore parameter uncertainty. A possible solution in future work might be the use of quantile–quantile plots[32], which take into account both the central estimates and the associated statistical distributions estimated by the model for each prediction. This permits an assessment of whether the model is overly confident or overly vague in its predictions, in addition to the accuracy of the central estimates.

There were some empirical shortcomings as well. Foremost, it was difficult to match the complex inclusion and exclusion criteria in two of the studies (CARDS and MICRO-HOPE), which created a potential difference between the actual patients and the hypothetical patients that mimic them. Further complicating matters, these were trials of anti-hypertensive and anti-dyslipidemia drugs and not of anti-diabetes drugs; as a diabetes model,

the central lever in ECHO-T2DM is the impact of changes in HbA1c on health outcomes.

Future validation work, moreover, could be improved by including more independent studies. We have already performed validation testing vs a number of newer randomized clinical trials (including ACCORD[33], ADVANCE[34], and ASPEN[35]) as part of the Mount Hood Challenges. The results are currently in press[13].

Finally, like all similar studies we are aware of, we were unable to identify published data on resource use and costs of T2DM suitable for testing the validity of the economic calculations. Given the importance of health economic end-points to estimates of cost-effectiveness (the general purpose of such modeling), this is an important limitation that should be addressed in future work.

Outlook

Work is ongoing to improve ECHO-T2DM's macrovascular risk engine and the new UKPDS risk equations are eagerly awaited.

Conclusions

Using currently available validation 'yardsticks', ECHO-T2DM simulated the treatment, progression, and patient outcomes observed in important clinical trials with accuracy consistent with other well-accepted models. We additionally extended previously used methodology by introducing a formal *F*-test of the scatterplot coefficients as a complement to the R^2 metric. Further development of a standardized methodology would improve the quality of health economic models by making validation techniques accessible to more practitioners and by increasing the comparability of results across studies.

Transparency

Declaration of funding
This work was financed by Janssen Global Services, LLC.

Declaration of financial/other relationships
JH is employed by Janssen Global Services, LLC. MW and CA work for IHE and ESiOR, respectively, both of which have provided consulting services for Janssen Global Services, LLC.

Acknowledgments
Cheryl Neslusan, PhD, of Janssen Global Services, LLC, and Pierre Johansen of the Swedish Institute for Health Economics have actively contributed to both the construction of ECHO-T2DM and to the preparation of this manuscript. Technical editorial assistance was provided by Kimberly Dittmar, PhD, of MedErgy and was funded by Janssen Global Services, LLC.

References

1. Drummond M, Schulper M, Torrance G, et al. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford: Oxford University Press; 2005
2. ADA. Guidelines for computer modeling of diabetes and its complications. Diabetes Care 2004;27:2262-5
3. Caro JJ, Briggs AH, Siebert U, et al. Modeling good research practices–Overview: A report of the ISPOR-SMDM modeling good research practices task force-1. Med Decis Making 2012;32:667-77
4. Willis M, Asseburg C, Borg S, et al. Cost-effectiveness of strict 'get to goal' treatment directives in the treatment of younger patients with newly diagnosed type 2 diabetes mellitus. 9th Annual ISPOR European Congress; September 18-19, 2010; Malmö, Sweden
5. Eastman RC, Javitt JC, Herman WH, et al. Model of complications of NIDDM. I. Model construction and assumptions. Diabetes Care 1997;20:725-34
6. Palmer AJ, Roze S, Valentine WJ, et al. The CORE diabetes model: Projecting long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making. Curr Med Res Opin 2004;20(Suppl 1):S5-26
7. Beale S, Bagust A, Shearer AT, et al. Cost-effectiveness of rosiglitazone combination therapy for the treatment of type 2 diabetes mellitus in the UK. Pharmacoeconomics 2006;24(Suppl 1):21-34
8. 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. Med Decis Making 2012;32:733-43
9. The Mount Hood 4 Modeling Group. Computer modeling of diabetes and its complications: A report on the fourth Mount Hood challenge meeting. Diabetes Care 2004;30:1638-46
10. Mount Hood Challenge homepage. Available at: https://sites.google.com/site/mounthoodchallenge/home [Last accessed 1 March 2013]
11. Brown JB, Palmer AJ, Bisgaard P, et al. The Mt. Hood challenge: cross-testing two diabetes simulation models. Diabetes Res Clin Pract 2000;50(Suppl 3):S57-64
12. Eastman RC, Javitt JC, Herman WH, et al. Model of complications of NIDDM. II. Analysis of the health benefits and cost-effectiveness of treating NIDDM with the goal of normoglycemia. Diabetes Care 1997;20:735-44
13. Palmer AJ, The Mount Hood 5 modeling group. Computer modeling of diabetes and its complications: A report on the fifth Mount Hood challenge meeting. Value Health 2013;In press.
14. 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
15. Eddy DM, Schlessinger L. Validation of the Archimedes diabetes model. Diabetes Care 2003;26:3102-10
16. Clarke PM, Gray AM, Briggs A, et al. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom prospective diabetes study (UKPDS) outcomes model (UKPDS no. 68). Diabetologia 2004;47:1747-59
17. UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998;352:837-53
18. Klein R, Klein BE, Moss SE, et al. Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy. JAMA 1988;260:2864-71
19. Thomason MJ, Colhoun HM, Livingstone SJ, et al. Baseline characteristics in the Collaborative AtoRvastatin Diabetes Study (CARDS) in patients with type 2 diabetes. Diabetic Med 2004;21:901-5
20. Colhoun HM, Betteridge DJ, Durrington PN, et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): Multicentre randomised placebo-controlled trial. Lancet 2004;364:685-96
21. Gerstein HC, Bosch J, Pogue J, et al. Rationale and design of a large study to evaluate the renal and cardiovascular effects of an ACE inhibitor and vitamin E in high-risk patients with diabetes. The MICRO-HOPE study.

Microalbuminuria, cardiovascular, and renal outcomes. Heart outcomes prevention evaluation. Diabetes Care 1996;19:1225-8
22. HOPE Study Investigators. Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: Results of the HOPE study and MICRO-HOPE substudy. Lancet 2000;355:253-9
23. Sasaki A, Uehara M, Horiuchi N, et al. A 15 year follow-up study of patients with non-insulin dependent diabetes mellitus (NIDDM) in Osaka, Japan. Long-term prognosis and causes of death. Diabetes Res Clin Pract 1996;34:47-55
24. Hoerger T, Segel J, Zhang P, et al. Validation of the CDC-RTI diabetes cost-effectiveness model. RTI Press publication No. MR-0013-0909. Research Triangle Park, NC: RTI International, 2009. Available at: http://www.rti.org/rtipress [Last accessed 1 March 2013]
25. Mueller E, Maxion-Bergemann S, Gultyaev D, et al. Development and validation of the Economic Assessment of Glycemic control and Long-term Effects of diabetes (EAGLE) model. Diabetes Technol Ther 2006;8:219-36
26. American Diabetes Association. Guidelines for computer modeling of diabetes and its complications. Diabetes Care 2004;27:2262-5
27. Simmons RK, Coleman RL, Price HC, et al. Performance of the UK prospective diabetes study risk engine and the Framingham risk equations in estimating cardiovascular disease in the EPIC- Norfolk cohort. Diabetes Care 2009;32:708-13
28. Kengne AP, Patel A, Colagiuri S, et al. The Framingham and UK Prospective Diabetes Study (UKPDS) risk equations do not reliably estimate the probability of cardiovascular events in a large ethnically diverse sample of patients with diabetes: The Action in Diabetes and Vascular Disease: Preterax and Diamicron-MR Controlled Evaluation (ADVANCE) study. Diabetologia 2010;53:821-31
29. van der Heijden AA, Ortegon MM, Niessen LW, et al. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. Diabetes Care 2009;32:2094-8
30. Zhou H, Isaman DJ, Messinger S, et al. A computer simulation model of diabetes progression, quality of life, and cost. Diabetes Care 2005;28:2856-63
31. Barhak J, Isaman DJ, Ye W, et al. Chronic disease modeling and simulation software. J Biomed Inform 2010;43:791-9
32. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. Biometrika 1968;55:1-17
33. Gerstein HC, Miller ME, Byington RP, et al. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 2008;358:2545-59
34. Patel A, MacMahon S, Chalmers J, et al. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. N Engl J Med 2008;358:2560-72
35. Knopp RH, d'Emden M, Smilde JG, et al. Efficacy and safety of atorvastatin in the prevention of cardiovascular end points in subjects with type 2 diabetes: the Atorvastatin Study for Prevention of Coronary Heart Disease Endpoints in non-insulin-dependent diabetes mellitus (ASPEN). Diabetes Care 2006;29:1478-85
36. Bagust A, Hopkinson PK, Maier W, et al. An economic model of the long-term health care burden of type II diabetes. Diabetologia 2001;44:2140-55
37. Brown JB, Russell A, Chan W, et al. The global diabetes model: User friendly version 3.0. Diabetes Res Clin Pract 2000;50(Suppl 3):S15-46
38. Bagust A, Evans M, Beale S, et al. A model of long-term metabolic progression of type 2 diabetes mellitus for evaluating treatment strategies. Pharmacoeconomics 2006;24(Suppl 1):5-19
39. Klein R, Klein BEK, Moss SE. Prevalence of microalbuminuria in older-onset diabetes. Diabetes Care 1993;16:1325-30
40. Dyck PJ, Kratz KM, Karnes, JL, et al. The prevalence by staged severity of various types of diabetic neuropathy, retinopathy, and nephropathy in a population-based cohort: the Rochester Diabetic Neuropathy Study. Neurology 1993;43:817-24
41. Palumbo, PJ, Melton SJ. Peripheral vascular disease and diabetes. In: Harris MI, Cowie CC, Stern MP, et al., editors. Diabetes in America. Bethesda: National Institutes of Health; 1995:401-408

