Validation of the Archimedes Diabetes Model

DAVID M. EDDY, MD, PHD
LEONARD SCHLESSINGER, PHD

OBJECTIVE — To validate the Archimedes model of diabetes and its complications for a variety of populations, organ systems, treatments, and outcomes.

RESEARCH DESIGN AND METHODS — We simulated a variety of randomized controlled trials by repeating in the model the steps taken for the real trials and comparing the results calculated by the model with the results of the trial. Eighteen trials were chosen by an independent advisory committee. Half the trials had been used to help build the model ("internal" or "dependent" validations); the other half had not. Those trials comprise "external" or "independent" validations.

RESULTS — A total of 74 validation exercises were conducted involving different treatments and outcomes in the 18 trials. For 71 of the 74 exercises there were no statistically significant differences between the results calculated by the model and the results observed in the trial. Considering only the trials that were never used to help build the model—the independent or external validations—the correlation was $r = 0.99$. Including all of the exercises, the correlation between the outcomes calculated by the model and the outcomes seen in the trials was $r = 0.99$. When the absolute differences in outcomes between the control and treatment groups were compared, the correlation coefficient was $r = 0.97$.

CONCLUSIONS — The Archimedes diabetes model is a realistic representation of the anatomy, pathophysiology, treatments, and outcomes pertinent to diabetes and its complications for applications that involve the populations, treatments, outcomes, and health care settings spanned by the trials.

Diabetes Care 26:3102–3110, 2003

The Archimedes diabetes model is described in a companion article in this issue (1). This article describes the validation of that model.

RESEARCH DESIGN AND METHODS

The purpose of any model is to estimate as accurately as possible for a given set of circumstances or actions whatever outcomes one wants to use the model to predict. To test how well the Archimedes diabetes model does this, we simulated a wide range of clinical trials. The studies were chosen by an independent advisory committee appointed by the American Diabetes Association. The trials were chosen by quality of design and importance of results and to collectively span a wide range of patient populations, organ systems, treatments, delivery settings, and outcomes. Half of the trials were used to help build the model ("internal" or "dependent" validations); the other half were not ("external" or "independent" validations).

For each validation exercise, we created a "virtual trial" by repeating the steps taken in the real trial and then compared the outcomes seen in the virtual trial with those that occurred in the real trial. To set up a validation exercise, we first had the model create a large virtual population containing a broad spectrum of ages, sexes, race/ethnicities, characteristics, behaviors, and diseases. We did this by having the model "give birth" to a very large number people of different sexes and race/ethnicities and letting them grow up (i.e., letting their physiologies function according to the equations described in the companion article). Information from the National Health and Nutrition Examination Survey (NHANES)-III on the marginal and joint distributions of patient characteristics and other risk factors is used to ensure that the population is representative of the U.S. population (2). Other populations could be constructed if desired (e.g., an Indian reservation).

In general, the steps we used to simulate a particular clinical trial were as follows. We began with the initial description of the trial, focusing in particular on the inclusion and exclusion criteria, treatment protocols, follow-up protocols, and definitions of the outcomes. We then had the model do the following. *1*) First, it searched the large population to identify people who met the entry criteria for the trial. Then it confirmed that their characteristics (e.g., age, sex, other conditions, treatments, and lab results) matched the distribution of characteristics published in the description of the trial. If not, over- or undersampling was performed as required, as would occur for a real trial. From that group, people were randomly selected to match the number of people in the trial. At the end of this selection process, the demographic, physiologic, and anatomic features, as well as the medical

From the Care Management Institute, Kaiser Permanente and Kaiser Permanente Southern California, Oakland, California.

Address correspondence and reprint requests to David M. Eddy, 1426 Crystal Lake Rd., Aspen, CO 81611. E-mail: eddyaspen@yahoo.com.

Received for publication 24 February 2003 and accepted in revised form 24 July 2003.

L.S. holds stock in Merck and Pfizer.

Abbreviations: 4S, Scandinavian Simvastatin Survival Study; CAD, coronary artery disease; CARE, Cholesterol and Recurrent Events; DCCT, Diabetes Control and Complications Trial; DPP, Diabetes Prevention Program; FPG, fasting plasma glucose; HHS, Helsinki Heart Study; HOPE, Health Outcomes Prevention Evaluation; HPS, Heart Protection Study; IDNT, Irbesartan Diabetic Nephropathy Trial; IRMA, Irbesartan in Patients with Type 2 Diabetes and Microalbuminuria; LIPID, Long-Term Intervention with Pravastatin in Ischemic Disease; LRC-CPPT, Lipid Research Clinics Coronary Primary Prevention Trial; MRC, Medical Research Council; SHEP, Systolic Hypertension in the Elderly Study; UKPDS, U.K. Prospective Diabetes Study; VA-HIT, Veterans Affairs High-Density Lipoprotein Cholesterol Interventions Trial; WOSCOPS, West of Scotland Coronary Prevention Study.

A table elsewhere in this issue shows conventional and Système International (SI) units and conversion factors for many substances.

See accompanying editorial, p. 3182.

histories of the people in the virtual trial, should match those of the people in the real trial, as far as can be determined from the publications and within sampling error. 2) If the description of the trial called for any interventions to be given before the people were randomized, such as a diet, then the simulated providers were instructed to give that intervention. 3) The people were then randomized into the number of groups used in the trial. 4) Simulated providers then gave the people in each group the designated treatments, using the protocols described for the trial. This included any important breaches in either provider or patient adherence that were described for the trial. 5) The people's physiologies were allowed to continue to function, including the effects of whatever treatments they were receiving, all as determined by the equations in the model. 6) Simulated providers then followed each patient with simulated appointments and tests, using the protocols and intervals described for the real trial. 7) In the model, as in the real trial, between scheduled visits patients could also develop symptoms, seek care, make appointments, have visits, be tested, be diagnosed, and be treated, all as determined by the equations. 8) The results were recorded at the time intervals used in the real trials. 9) The results were then processed and compared with those described for the real trial.

All of this was done at whatever level of detail was necessary to simulate what was done in the real trial, using whatever descriptions were available from the publications. For example, if two trials reported retinopathy outcomes but one measured two-step retinopathy (3), whereas the other measured three-step retinopathy (4), we had the simulated physicians apply the appropriate protocol to the appropriate trial. This also applies to inclusion criteria. If hypertension was defined as "a finding on at least two of three consecutive measurements obtained 1 week apart. . . of a mean systolic blood pressure >135 mmHg or mean diastolic blood pressure >85 mmHg, or both" (5), then these were the guidelines that we had the simulated physicians follow. When the description of a trial included variables or diseases that were not yet in the Archimedes model, we ignored them. For example, the model does not yet include pregnancy. If a trial excluded pregnant women, we ignored that exclusion criterion. If a trial included variables or conditions that were not yet in the model at the time the simulation was requested, we expanded the model to include those factors before performing the simulation. If any information from such a trial was used to help expand the model, we noted that fact and classified the resulting validation as an internal or dependent validation. (The use of trial information will be described more in detail below.) For example, before the Irbesartan in Patients with Type 2 Diabetes and Microalbuminuria 2 trial (IRMA) (5) could be simulated, we had to expand the part of the model that represented the progression of untreated nephropathy at high levels of albuminuria and the effects of angiotensin-II receptor antagonists on glomerular function. The IRMA trial is therefore considered an internal or dependent validation.

The trials

The model was validated against 18 trials, all chosen by the independent advisory committee. Ten trials explicitly included people with diabetes. These are the U.K. Prospective Diabetes Study (UKPDS) (3), the Diabetes Prevention Program (DPP) (6), the Heart Protection Study (HPS) (7), the Health Outcomes Prevention Evaluation (HOPE) (8), Micro-HOPE (the diabetic subpopulation of the HOPE trial) (9), Cholesterol and Recurrent Events (CARE) (10), the ACE Inhibitors and Diabetic Nephropathy Trial (Lewis) (11), the IRMA-2 trial (5), the Diabetes Control and Complications Trial (DCCT) (4), and the Irbesartan Diabetic Nephropathy Trial (IDNT) (12). The CARE trial has also published results for age-group subpopulations (13). Eight more trials were chosen by the committee to test the model's realism for representing coronary artery disease (CAD). They are the Long-Term Intervention with Pravastatin in Ischemic Disease (LIPID) trial (14), the Helsinki Heart Study (HHS) (15), the Systolic Hypertension in the Elderly Study (SHEP) (16), the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) (17), the Medical Research Council (MRC) hypertension trial (18), the West of Scotland Coronary Prevention Study (WOSCOPS) (19), the Veterans Affairs High-Density Lipoprotein Cholesterol Interventions Trial (VA-HIT) (20), and the Scandinavian Simvastatin Survival Study (4S) (21).

Use of trial data to build the model. Ten of the trials (DPP, HPS, MICRO-HOPE, LIPID, HHS, SHEP, LRC-CPPT, MRC, VA-HIT, and WOSCOPS) were not used at all to build the physiology model; they provided external or independent validations of the model. The remaining eight trials (UKPDS, HOPE, CARE, Lewis, IRMA-2, DCCT, IDNT, and 4-S) provided internal or dependent validations. For these, the type of use varied from trial to trial but can be summarized as follows. In general, between 10 and 30 equations are needed to represent the pathophysiology of the disease and to calculate the effect of a specific treatment on a specific outcome in a specific population (i.e., not including the equations for behaviors, care processes, logistics, and other nonbiological aspects of the model). When a piece of information from a trial is used, it is used to help write only one of those 10–30 equations. A trial that significantly pushes the boundaries of the model might contribute two or three pieces of information, each to a particular equation. A trial's results are never used to write or "fit" an equation, such as a regression equation or transition probability, that directly relates the population, treatment, and outcome. Indeed, there are no such equations in the model. When the results being matched are sampled outcomes, an iterative method is used, stopping when the calculated result and real result are within ±1 SD.

When information from a trial is used to help build the model, it is used to build some new or deeper part of the model. Thereafter, that part is used in all subsequent simulations. For example, the 4S trial was the primary source for information about the possible direct effects of Simvastatin on rates of coronary artery occlusion. Thereafter, that equation was used for all subsequent simulations involving statins; for example, simulation of the HPS study of Simvastatin did not use any data from the HPS trial. The high accuracy of the simulation of the HPS (Table 1) provides an independent check on the equation fitted with 4S data. As each new equation is written, it becomes a permanent part of the model; the parameters of an equation are never changed to fit particular trials. Previously performed simulations are rerun as needed to ensure that as the model advances it remains accurate for all of the trials. At any time there is always a single set of equations, and those

Table 1—*Comparison of model and trial results: trials that include people with diabetes*

Name of trial	Population	Outcome	Years	Initial size	Treatment group	Result (%)	
						Model	Trial
UKPDS	Newly diagnosed type 2 diabetes	Myocardial infarction	12	1,138	Conventional	19.6	19
				2,729	Intensive*	15.4	16
		Albuminuria	12	1,138	Conventional	33.8	34
				2,729	Intensive	21.3	23
		Proteinuria	12	1,138	Conventional	9.8	10.3
				2,729	Intensive	7.6	6.8
		Retinopathy	12	1,138	Conventional	50	49
				2,729	Intensive	39	39
DPP†	Impaired glucose tolerance, Impaired fasting glucose and Overweight	Progression to diabetes	4	1,082	Control	38	37
				1,073	Metformin	31	28
				1,079	Lifestyle	21	20
HPS†	High risk for CAD events‡	Major coronary events	5	10,267	Placebo	11.7	11.8
				10,269	Simvastatin	8	8.8
		CHD death	5	10,267	Placebo	6.2	6.9
				10,269	Simvastatin	5	5.5
HOPE	High CAD risk§	Myocardial infarction	4.5	4,652	Placebo	11.3	11.3
				4,645	Ramipril	8.9	9
MICRO-HOPE†	High CAD risk, type 2 diabetes	Myocardial infarction	4	1,808	Placebo	13	12.9
				1,769	Ramipril	9	10.2
CARE‖	Recent myocardial infarction, average cholesterol	Myocardial infarction	5	2,078	Placebo	12.3	13.2
				2,081	Simvastatin	9.3	10.2
		CHD death	5	2,078	Placebo	6.2	5.7
				2,081	Simvastatin	4.4	4.6
Lewis	Type 1 diabetes, nephropathy	Doubling of creatinine	4	202	Placebo	37	33
				207	Captopril	19	22
IRMA-2	Type 2 diabetes, micro-albuminurea	Nephropathy	1.8	201	Placebo	17.4	15
				195	Irbesartan 150	9.5	9
				194	Irbesartan 300	5.3	4.5
DCCT primary	Type 1 diabetes without retinopathy	Retinopathy	8	378	Loose control	34	38
				348	Tight control	9.3	10
		Albuminuria	8	378	Loose control	29	28
				348	Tight control	17	15
		Proteinuria	9	378	Loose control	32	25
				348	Tight control	15	18
DCCT secondary	Type 1 diabetes with retinopathy	Retinopathy	8	352	Loose control	52	48
				363	Tight control	22	21
		Albuminuria	8	352	Loose control	33	35
				363	Tight control	22	22
		Proteinuria	9	352	Loose control	9	11
				363	Tight control	5	6
IDNT	Type 2 diabetes, nephropathy	Doubling of creatinine	4	579	Placebo	35	37
				569	Irbesartan	26	28

*Sulphonylurea, Metformin, or insulin; †not used to build physiology model; ‡CAD, occlusive arterial disease or diabetes; §CAD or diabetes plus at least one CVD risk factor; ‖eight additional validation exercises were done for the under-60 and over-60 age-groups. No model results were significantly different from trial results.

equations reproduce or predict every trial. The fact that the model is anchored to such a wide variety of populations, treatments, and outcomes guards against overspecification of the model.

With that background, the actual uses of trial data were as follows. Two trials contributed to the model of glucose homeostasis and the development and progression of diabetes (Fig. 1 of the companion article [1]). Specifically, the average fasting plasma glucose (FPG) in the control group of the UKPDS trial (22) was used to help write an equation for the effects of insulin resistance, and the DCCT's results were used to help model the development of type 1 diabetes and the effect of glucose control. Data from several trials were used to help build models of the complications of diabetes. We used data from the UKPDS to help write the equations for the retinopathy and nephropathy features. The DCCT was used to help model the progression of nephropathy and retinopathy in patients with type 1 diabetes. The HOPE trial was used to model the effects of ACE inhibi-

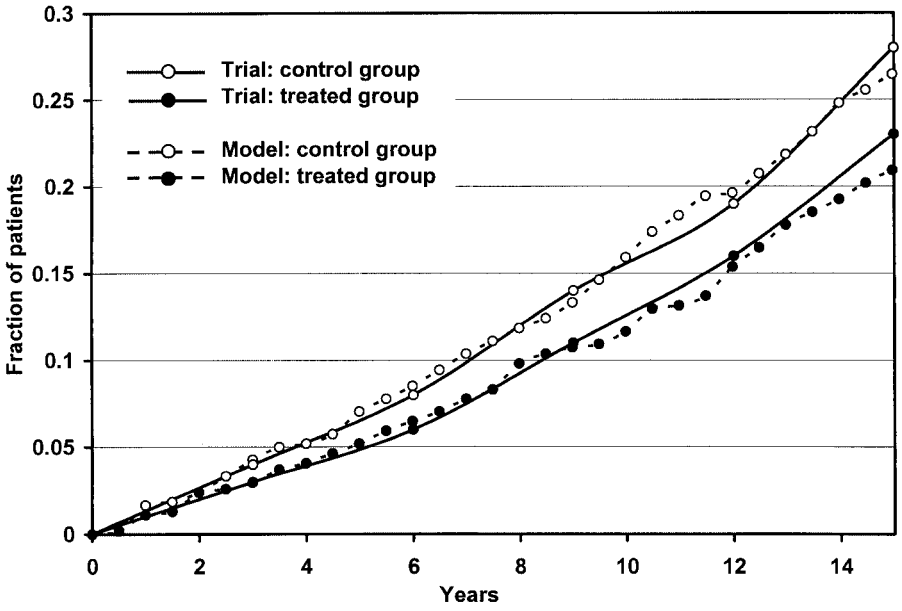

Figure 1—*Comparison of model and trial: fraction of patients having myocardial infarctions in the UKPDS.*

tors on variables such as peripheral resistance, fast and slow occlusion, the action of thrombolytics, and the progression of congestive heart failure. Data from the CARE trial were used to help build the part of the model that determines survival following myocardial infarction as a function of the proportion of the myocardium affected by myocardial infarction and the recovery of the myocardium following nonfatal myocardial infarction. Data from the Lewis trial were used to help estimate the progression of glomerular damage in people with established and severe nephropathy. Information from the IRMA-2 trial was used to model the progression of untreated nephropathy at high levels of albuminuria and to model the effects of angiotensin-II receptor antagonists on glomerular function. Finally, data from the 4S trial were used to model the effect of statins on development of thrombi.

Goodness of fit. To determine the accuracy of the model, we focused on outcomes determined by the underlying disease. Outcomes that are likely to be heavily influenced by local practices, such as the rate of bypasses, or by nondiabetes factors, such as deaths from other causes, have questionable external validity and were not included.

For the disease-determined outcomes, we use Kaplan-Meier curves to compare the results calculated by the model with the actual results of the trial. Kaplan-Meier curves provide the most complete information about the outcomes over the entire time course of the trial in all the arms of a trial. Because the results of both a real trial and the model are subject to random variation, one would not expect the Kaplan-Meier curves to match exactly. Our approach is to calculate whether the differences are statistically significant or could be explained by chance. The published reports of trials rarely contain sufficient information to perform precise statistical comparisons of Kaplan-Meier curves. Specifically, because entry into a trial is usually staggered, the number of people actually followed to the last reported year of a trial is usually much smaller than the number of people entered, typically <25% of the starting sample size. To calculate the statistical significances of the differences, we used a very conservative method that assumes that everyone entered into a trial is followed for its full duration, with the provision that if there are known to be <100 people at the last follow-up time, we would use the results from the previous observation period. This method biases against the model because it greatly underestimates the random variation that affects the results toward the end of the real trial. With that limitation, we determine for each arm of a trial whether the difference between the trial and the model are statistically significant at the $P = 0.05$ level (corrected χ^2). If not, we say that the model's results "statistically match" the trial's results. To gain an overview of the complete body of validation exercises, we also calculated cor-

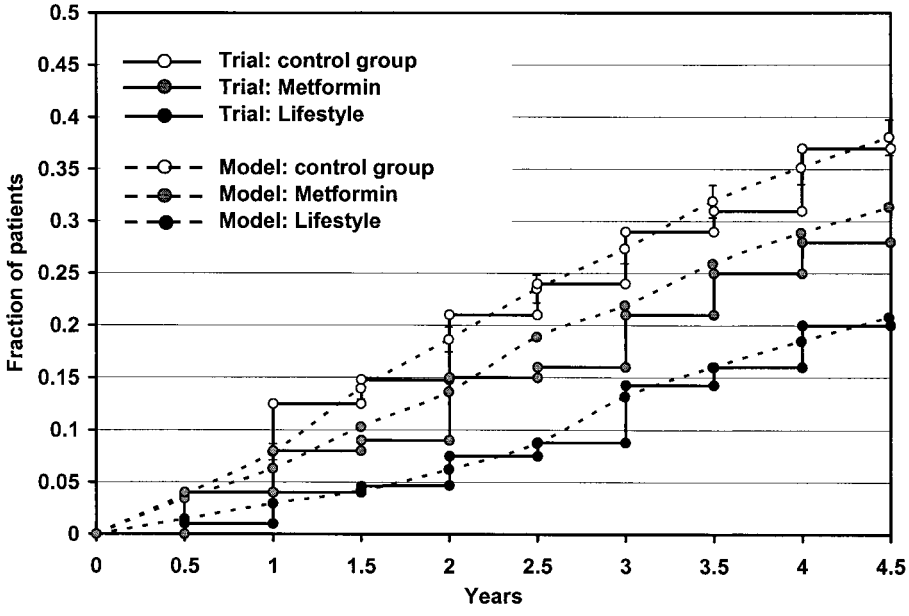

Figure 2—*Comparison of model and trial: Fraction of patients developing diabetes in the Diabetes Prevention Program*

Table 2—*Comparison of model and trial results: for trials of CAD*

Name of trial	Population	Outcome	Years	Initial size	Treatment group	Result (%) Model	Result (%) Trial
LIPID*	Acute MI within 3–36 months, "broad range" of lipid levels	CHD death	6	4,502	Placebo	7.5	8.3
				4,512	Pravastatin	6.5	6
		Myocardial infarction	6	4,502	Placebo	14.4	15.6
				4,512	Pravastatin	11	12
HHS*	Middle aged men, dyslipidemia	Myocardial infarction	5	2,030	Placebo	4.2	4.1
				2,051	Gemfibrozil	3	2.7
4S	History of angina or acute myocardial infarction	Myocardial infarction	5.4	2,223	Placebo	23.8	25
				2,221	Simvastatin	14.2	16
SHEP*	Isolated systolic hypertension	CAD events	4.5	2,371	Placebo	5.8	5.9
				2,365	Antihypertensive†	4.5	4.3
LRC*	Primary hypercholesterolemia	Myocardial infarction	4.5	1,543	Placebo	5.4	6
				1,543	Cholestyramine	4	5
MRC*	Mild hypertension	Myocardial infarctions	4	8,677	Placebo	4.5	4.5
				8,677	Antihypertensive‡	3.3	3.4
WOSCOPS*	Very high risk and hypercholesterolemia	Myocardial infarctions	5	3,293	Placebo	5.2	7.9§
				3,302	Pravastatin	2.6	5
		Coronary heart disease deaths	5	2,078	Placebo	1.9	1.7
				2,081	Pravastatin	1.1	1.2
VA-HIT	Previous CAD, low HDL	Myocardial infarctions	5	1,264	Placebo	25.2	23
				1,267	Gemfibrozil	17.8	19.7
		Coronary heart disease death	5	1,264	Placebo	10.2	9.6
				1,267	Gemfibrozil	8.7	8.4
		Stroke	5	1,264	Placebo	4.2	6.6
				1,267	Gemfibrozil	3.5	5.2

*Not used to build physiology model; †step 1: Chlorthalidone, step 2: Atenolol; ‡Bendrofluazide or propranolol; §difference between model results and trial results statistically significant, $P < 0.01$; ‖difference between model results and trial results statistically significant, $P < 0.05$.

relation coefficients for the two sets of results.

RESULTS

Including each arm and each outcome reported in a trial as a validation exercise to date, the model has been subjected to 74 validation exercises involving the 18 trials. The use of Kaplan-Meier curves to compare the results of the model and trial are illustrated in Figs. 1 and 2. Figure 1 shows the curves calculated by the model and reported for the trial for the fraction of people who develop fatal or nonfatal myocardial infarctions in the UKPDS (3), a trial that was used to help build the model and thus represents an internal or dependent validation. This also illustrates the relatively unstable results that can occur at the longest follow-up time due to the steady decrease in sample size over time; <100 patients were followed for the full 15 years in the UKPDS. Figure 2 shows the results for the DPP. The DPP was not used to build the model; the results of the model were calculated based on the initial descriptions of the trial and publicly presented before the actual outcomes of the trial were published.

The results for the 10 trials that explicitly included patients with diabetes are summarized in Table 1. The results of the other trials that are pertinent to the cardiovascular complications of diabetes are summarized in Table 2. Trials not used to build the model are marked.

Goodness of fit

Of 74 validation exercises, the results of the model statistically matched the results of the trial in all but three exercises. Each of these was from a trial that was not used to build the model. In one of them, the stroke outcome in the placebo group of the VA-HIT trial, the results just barely reached statistical significance ($P = 0.04$), which is to be expected in 74 exercises. For this exercise, the model still estimated the effect of the treatment with good accuracy (1.7 vs. 1.4%, $P > 0.05$). The only exercises that showed a highly significant difference between the results of the model and the trial came from the WOSCOPS trial (19), which is discussed further. The correlation coefficient for all 74 exercises is $r = 0.99$ (Fig. 3). If the outcomes in the control group and the absolute differences between the control and treated groups are compared for model and trial, the correlation coefficient is $r = 0.99$. Focusing specifically on the absolute differences in the outcomes, which determines the number needed to treat, the correlation coefficient is $r = 0.97$. For the 10 trials that were not used to build the model, the correlation coefficient is also $r = 0.99$. (This includes the three discrepant results.)

CONCLUSIONS — Our objective for the Archimedes diabetes model is to create a "virtual world" that represents clinical reality as realistically as is reasonably possible given today's information and modeling methods. Once created, the model can be used to address a variety of clinical problems and questions. For example, interventions, guidelines, performance measures, disease management programs, strategic goals, implementation strategies, continuous quality im-

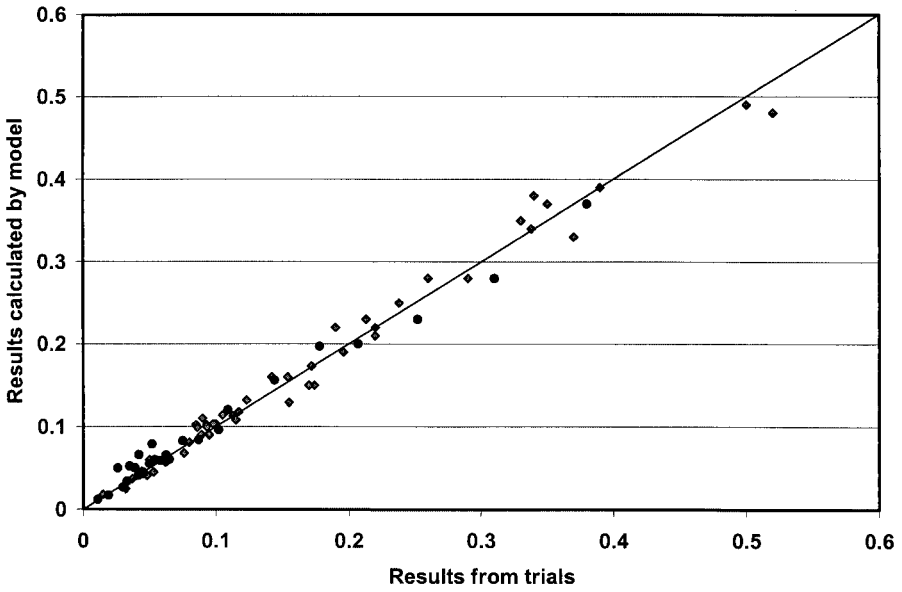

Figure 3—*Comparison of the results calculated by model with the results of the actual trials for 74 validation exercises. Filled circles compare the results calculated by the trials (x-axis) and the results calculated by the model (y-axis) for independent or external validation exercises. Gray diamonds compare the results for dependent or internal validation exercises. The 45° line indicates perfect accuracy. The results will deviate from this line due to random factors as well as any inaccuracies in the model.*

provement projects, and research projects can be "tried out" and optimized in the virtual world of the model in ways that may not be feasible in the real world.

The ability to use a model for these purposes depends critically on the accuracy of its estimates. Ultimately, this requires comparisons to real experiences. A starting point is to use the model to simulate real clinical trials and compare the results. If the results match within the expected degree of sampling variation, we gain confidence that the model's representation of the pathophysiology of diabetes and its complications is reasonably realistic. This in turn builds confidence that the results of future applications will be reasonably accurate, at least for the populations, organ systems, treatments, outcomes, and care settings represented by the trials. To our knowledge, no other model in health care—for diabetes, CAD, or any other condition—has been validated against clinical trials as we describe here (23).

The results obtained to date are encouraging. First, they do suggest that the model is reasonably realistic. But they also carry information concerning the body of knowledge about diabetes and CAD that has been built up over the years. There are many remaining uncertainties and controversies about the pathogenesis of diabetes. These validations provide some assurance that despite these gaps, the information that does exist, at least as interpreted through this model, provides a reasonably sound basis for making decisions and setting policies.

Perhaps the single most important feature that distinguishes the Archimedes model from other clinical models is that it is based on a representation of human biology. The primary motivation for taking this approach is that it is the only way to achieve the objectives we had for the model, as described elsewhere (1). This justification notwithstanding, it is still reasonable to ask whether a nonbiological model might be able to achieve the same degree of predictive accuracy. We believe that it would be extremely difficult. The immediate problem is that just performing the validations requires a high degree of biological detail. For example, the outcome of the IRMA-2 trial was "a urinary albumin excretion rate that was >200 μg/min and at least 30% higher than the baseline rate on at least two consecutive visits" (5). To simulate that trial, a model would need to estimate the effect of Irbesartan on that outcome, as determined by that protocol, in 590 patients who ranged in age from 30 to 70 years, had type 2 diabetes of various durations, had systolic blood pressures >135 mmHg and diastolic blood pressures >85 mmHg, had initial albumin excretion rates ranging from 20 to 200 μg/min, had serum creatinine concentrations not exceeding 1.5 mg/dl, etc. Then, to deliver the same results as Archimedes, the model would need to simulate 17 other trials, each of which has an equally complex list of biological variables that address other populations, treatments, and outcomes. It is difficult to imagine how this would be done without a robust model of biology at the level of detail defined by these variables. Some may question the need for this level of detail in a simulation, but our operating principle is that if researchers and clinicians consider a variable or process sufficiently critical to be made part of a trial's protocol, then it should be considered critical in the simulation of that protocol. Even if testing protocols can be loosened (e.g., "the average of three readings a week apart"), certainly the inclusion criteria (e.g., "an myocardial infarction in the past 2–30 months"), and outcomes (e.g., "increase in urinary protein of 30% over baseline") are critical.

There is very little empirical evidence to bring to this question because there are extremely few validations of other models against clinical trials using any methodology. A complete analysis of this literature is beyond this article, but cautionary flags are raised by findings that the Framingham equation, which is the core of most Markov models of CAD complications in diabetes, was "disappointing" because of its inability to predict the incidence of CAD events in the Cardiff Diabetes Registry (24), although firm conclusions cannot be drawn due to the methods used in that study. The Framingham equation also misestimated CAD events in the UKPDS, by a factor of almost two for CAD events and a factor of five for coronary heart disease mortality (25). A comparison of the UKPDS risk engine and the Joint British Society (JBS) method found highly significant and clinically important differences in the proportions of people classified into different risk groups by those two models (26). Estimates by the Framingham and Prospective Cardiovascular Munster (PROCAM) models of the risk of coronary heart disease events in people with diabetes varied by a factor of more than two (27). Also noteworthy are the findings of wide differences in cross

tests of different diabetes models, even when each model is handed identical patients (28,29).

It is important to stress several points about the validation exercises. First, each of these exercises involves a very deep simulation. In each, the predicted results come from thousands of simulated individuals. Each of them has a simulated liver, heart, pancreas, and other organs. Each liver is producing glucose, each coronary artery can develop plaque or thrombus at any point in any artery, each kidney is clearing urine, and so forth. All told, each simulation involves scores of equations in every patient; they all have to work together correctly over long periods of simulated time in order to generate the outcomes seen in the virtual trials. In general, the results for the control groups test the realism of the model's representation of the natural history of the disease, and the effects of the risk factors, patient characteristics, previous medical histories, severity of disease, and previous and concurrent medications, as described in the designs of the trials. The results for the treated groups test the model for all these plus the effects of the treatments.

A second point is that together these validations crisscross virtually every aspect of diabetes and its complications (see Table 1). We believe a model should be considered "validated" only for applications that are spanned by the trials used to validate it. A measure of this is whether the populations, treatments and outcomes for a proposed application have each been included in at least one trial against which the model has been validated. Thus the multiple validations reported here are not redundant; each is probing different parts of the model.

Third, whenever data from a trial were used to help build the model (i.e., the internal or dependent validations), they were used to address a very specific aspect of the underlying physiological process, usually one equation out of dozens that are needed to complete a calculation. For example, the rate of increase of FPG in the "conventional policy" group of the UKPDS was used to help build the model. But it was not used to fit an equation for FPG. It was used to help write the equation that describes the effect of insulin resistance on hepatic glucose production and the uptake of glucose by muscle (Equation 10 in the companion article). Nine other equations are also needed to simulate glucose homeostasis and calculate a person's FPG, and many more equations are needed to calculate an end point like a myocardial infarction. Thus even when a validation exercise for Archimedes involves a trial that contributed some information to the model, in order for the validation to be successful dozens of other equations that were not touched by the trial need to function correctly. The validations of the eight trials that contributed to the model can be considered not only confirmations of the particular equations that were affected by each particular trial, but also independent tests of all the other equations needed to complete the calculations. Furthermore, an equation that was touched by any particular trial was independently validated by all the other exercises that involved other trials.

The validations have several limitations. First, the fact that they demonstrate the realism of our representation of the underlying biology of the disease does not mean that our representation is the only one capable of accomplishing similarly accurate predictions. All that can be said now is that the representation we have chosen is successful in producing accurate results for a wide variety of populations, treatments, outcomes, and settings. We know of no other representations that have been tested in this way that would permit any comparisons.

Second, the validations indicate that the model simulates what happens in clinical trials. However, this does not necessarily document the model's accuracy for predicting what happens outside of trials. The issue is "efficacy" vs. "effectiveness." The fact that patient and physician behaviors may be different in research settings than nonresearch settings affects all approaches for interpreting clinical trials, including expert judgment. The barrier to conducting validations outside of research settings is the availability of the necessary data. On the positive side, Archimedes includes the features needed to perform such validations, such as patient and practitioner behaviors, failure to follow protocols or reach treatment goals, both random and systematic variations in practices, errors in conducting or interpreting tests, and so forth. As better information on these factors becomes available from computerized medical records, we will perform these types of "effectiveness" validations. In the meantime, the model can be used to explore the potential effects of these factors and to identify which aspects of care processes are most important to monitor.

A similar limitation concerns the loss of patients to follow-up in real trials. In real trials, patients who die or are lost to follow-up are censored in the calculation of Kaplan-Meier curves. In the model, censoring can occur due to deaths. However, we do not model other reasons for losing patients unless the necessary information is published. This has the implication of assuming that, in a real trial, there are no patient selection biases affecting which patients are lost to follow-up. If there were information on this, Archimedes could include it. Lacking that, the model can be used to explore the potential importance of such biases.

A fourth limitation is that the model does not attempt to represent the underlying biology for causes of death other than diabetes and its complications, CAD, congestive heart failure, and asthma. The validation exercises presented here only address causes of death related to diabetes and CAD.

Fifth, our validation methods ignore variables or conditions (e.g., pregnant women) that might have been in the exclusion criteria of a trial but that are not in the model. In essence, this means that the validations are testing the explanatory power of the variables and conditions that the model does currently include.

A sixth limitation of the validations is that they do not evaluate any care processes that go beyond those that are described as part of a trial's protocol. The validations also do not address the logistics, resources, or costs involved in the model. These factors can vary from setting to setting and cannot be validated in any general sense, the way a representation of human physiology or the effect of a treatment can. Our approach to this is to enable users to check the care processes and resources that are currently in the model and modify them as needed.

A seventh limitation derives from the fact that the model has been validated against the average or aggregated outcomes for populations because that is the information available from the published trials. The simulations do reproduce the complex spectrum of ages, race/ethnicities, previous medical histories, and so forth that are in each trial, but the outcomes have to be averaged before they can be compared with the averages published for

the trials. Some trials, such as the CARE and HOPE trials, have published results for some subpopulations, and Archimedes matches those results very well. Furthermore, the wide mix of populations and other factors across the different trials provides a between-trial check on the model's realism for those factors—the model delivers accurate average results no matter how the populations and factors are mixed in the different trials. However, a more systematic analysis of these issues requires patient-specific information from trials.

Regarding individual patients, there are theoretical limits on the extent to which any model can ever be validated for predicting the outcomes for a particular patient. All we can say about the Archimedes model from these validations is that it has been reasonably accurate for a wide spectrum of populations with different mixtures of ages, sexes, race/ethnicities, complications, severities of disease, prior histories, concurrent treatments, and comorbid conditions. When person-specific data from clinical information systems become available, that potential use of the model can be explored in greater depth.

When a mismatch in a validation occurs, we examine it to determine its cause and whether any revisions to the model are appropriate. In the 74 validation exercises conducted thus far, the results for only one trial were substantially different from the real results. The discrepancy occurred in the control group of the WOSCOPS trial, where the model underestimated the rate of CAD events by ~35%. The model still predicted the absolute effect of Pravastatin accurately. The discrepancy in the background rate of CAD events could be due to the presence of a risk factor in that population that was not measured or reported in the description of the trial and therefore could not be included in the trial. Alternatively, it could be that the model's representation of physiology is not accurate for that particular population. We have not added a "WOSCOPS factor" to the model to make it match this trial's results.

This example emphasizes the fact that failure of a validation exercise does not necessarily mean the model is flawed. In addition to discrepancies due to random variations, the results can be thrown off by any changes in the treatment protocols in the real trial that are not described, poor adherence to protocols or treatments by practitioners or patients, incomplete follow-up, and/or important facts about the population that are not completely understood or described by the investigators. But beyond these is the fact that the intervention being studied in the trial might contain some surprises. Indeed, that is the very reason most trials are done. When a new trial reveals a result that could not have been predicted, we rejoice with everyone else about learning the new information and use it to advance the model.

Conversely, successful prediction of a trial's results by the model without any use of the trial's data, as occurred here for more than half of the trials (6,7,9,14–20), does not mean that these trials should not have been done. For example, the DPP not only confirms the interpretation of the previous research, which is very important in its own right, but also suggests that there are no surprises; our current understanding of the early natural history of the disease (at least as described by this model) appears to be correct. Furthermore, the DPP collected patient-specific data, which if analyzed with methods we describe elsewhere (30), could greatly increase our understanding of the pathophysiology of the disease.

In summary, we have tried to build a model that operates at the level of detail that clinicians and administrators consider important for their decisions. To strengthen the credibility and usefulness of the model we have tested it against a wide range of clinical trials. It appears to be a good representation of the anatomy, pathophysiology, tests, treatments, and outcomes of diabetes and its complications for applications that involve the range of populations, organ systems, treatments, outcomes, and care processes spanned by these trials. As additional information becomes available the model will be expanded and revalidated as needed.

Acknowledgments— The order of authorship is alphabetical. The development of this model was supported by Kaiser Permanente Southern California and the Care Management Institute of Kaiser Permanente. The advisory committee and performance of the validation exercises was supported in part by Bristol Myers Squibb through an educational grant to the American Diabetes Association.

We thank the members of the advisory committee, especially Richard Kahn and John Buse, for overseeing the validations.

References

1. Eddy DM, Schlessinger L: Archimedes: a trial-validated model of diabetes. *Diabetes Care* 26:3093–3101, 2003
2. Third National Health and Nutrition Examination Survey (NHANES III, 1994) CD ROM Series 11, No 1. Hyattsville, MD, National Center for Health Statistics, 1988
3. 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* 352:837–852, 1998
4. The Diabetes Control and Complications Trial Research Group: The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 329:977–986, 1993
5. Irbesartin in Patients With Type-2 Diabetes and Microalbuminuria Study Group: The effect of irbesartan on the development of diabetic nephropathy in patients with type 2 diabetes. *N Engl J Med* 345:870–878, 2001
6. Diabetes Prevention Program Research Group: Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med* 356:393–402, 2002
7. Heart Protection Study Collaborative Group: MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20,536 high-risk individuals: a randomized placebo-controlled trial. *Lancet* 360:23–33, 2002
8. The Heart Outcomes Prevention Evaluation Study Investigators: Effects of an angiotension-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *N Engl J Med* 342:145–153, 2000
9. The Heart Outcomes Prevention Evaluation 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* 355:253–259, 2000
10. Sacks FM, Pfeffer MA, Moye LA, Rouleau JL, Rutherford JD, Cole TG, Brown L, Warnica JW, Arnold JMO, Wun CC, Davis BR, Braunwald E: The effect of Pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *N Engl J Med* 335: 1001–1009, 1996
11. Lewis EJ, Hunsicker LG, Clarke WR,

Bain, Raymond P, Berl T, Rohde R, Raz I: The effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy. *N Engl J Med* 329:1456–1462, 1993
12. Lewis EJ, Hunsicker LG, Clarke WR, Tomas P, Pohl MA, Lewis JB, Ritz E, Atkins RC, Rohde R, Raz I: Renoprotective effect of the angioitensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N Engl J Med* 345: 851–860, 2001
13. Lewis SJ, Moye LA, Sacks FM, Johnstone DE, Timmis G, Mitchell J, Limacher M, Kell S, Glasser SP, Grant J, Davis Barry R, Pfeffer MA, Braunwald E: Effect of pravastatin on cardiovascular events in older patients with myocardial infarction and cholesterol levels in the average range: results of the cholesterol and recurrent events (CARE) trial. *Ann Intern Med* 129: 681–689, 1998
14. LIPID Study Group: Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. *N Engl J Med* 339:1349–1357, 1998
15. Frick MH, EloO, Haapa K, Heinonen OP, Heinsalme P, Helo P, Huttunen JK, Kaitaniemi P. Koskinen P, Manninen X, et al: Helsinki Heart Study: primary-prevention trial with gemfibrozil in middle-aged men with dyslipidemia: safety of treatment, changes in risk factors, and incidence of coronary heart disease. *N Engl J Med* 317:1237–1245, 1987
16. SHEP Cooperative Research Group: Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA* 265:3255–3264, 1991
17. The Lipid Research Clinics Coronary Primary Prevention Trial results. I. Reduction in incidence of coronary heart disease. *JAMA* 251:351–364, 1984
18. Medical Research Council Working Party: MRC trial of treatment of mild hypertension: principal results. *BMJ* 322:97–104, 2001
19. Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, Macfarlane PW, McKillop JH, Packard CJ, the West of Scotland Coronary Prevention Study Group: Prevention of coronary heart disease with Pravastatin in men with hypercholesterolemia. *N Engl J Med* 333:1301–1307, 1995
20. Rubins HB, Robins SJ, Collins D, Fye CL, Anderson JW, Elam MB, Faas FH, Linares E, Schaefer EJ, Schectman G, Wilt TJ, Wittes J: Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. *N Engl J Med* 341: 410–418, 1999
21. Scandinavian Simvastatin Survival Study Group: Randomized trial of cholesterol in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 344:1383–1389, 1994
22. UK Prospective Diabetes Study Group: Relative efficacy of randomly allocated diet, sulphonylurea, insulin, or Metformin in patients with newly diagnosed non-insulin dependent diabetes followed for three years (UKPDS 13). *BMJ* 310:83–88, 1995
23. Medline search for "model" and "validation," "validity," or "accuracy." Available from http://medlineplus.gov./ Accessed 19 September 2003
24. McEwan PC, Peters J, Currie CJ, Hopkins P, Griffiths JD, Williams JE: The unreliability of Framingham risk equations in predicting coronary heart disease (CHD) events in diabetes (Abstract). *Diabetes* 49 (Suppl. 1):A187, 2000
25. Yeo WW, Yeo KR: Predicting CHD risk in patients with diabetes mellitus. *Diabet Med* 18:341–344, 2001
26. Song S, Brown P: Comparison of UKPDS Risk Engine Model with Framingham-Based Method in the assessment of CHD risk in patients with diabetes mellitus and its clinical implications (Presented at ADA Annual Meeting, New Orleans, LA, 14 June 2003). *Diabetes* 52 (Suppl. 1):41-OR, 2003
27. Game FL, Jones AF: Coronary heart disease risk assessment in diabetes mellitus: a comparison of PROCAM and Framingham risk assessment functions. Diabetes UK. *Diabet Med* 18:355–359, 2001
28. Brown JB, Palmer AJ Bisgaard P, han W, Pedula K, Russell A: The Mt. Hood Challenge: cross-testing two diabetes simulations models. *Diabetes Res Clin Pract* 50 (Suppl.):S57–S64, 2000
29. Mt. Hood Challenge. II. San Francisco, CA, 12–13 June 2002
30. Schlessinger L, Eddy DM: Archimedes: a new model for simulating health care systems: the mathematical formulation. *J Biomedical Informatics* 35:37–50, 2002