SOLUCIA, INC.

Introduction to Predictive Modeling

December 13, 2007

Introduction / Objective

SOLUCIA, INC.

1. What is Predictive Modeling?
2. Types of predictive models.
3. Applications – case studies.

Predictive Modeling:
A Review of the Basics

SOLUCIA, INC.

Definition of Predictive Modeling

- ***"Predictive modeling is a set of tools used to stratify a population according to its risk of nearly any outcome…ideally, patients are risk-stratified to identify opportunities for intervention <u>before</u> the occurrence of adverse outcomes that result in increased medical costs."***

Cousins MS, Shickle LM, Bander JA. An introduction to predictive modeling for disease management risk stratification. Disease Management 2002;5:157-167.

SOLUCIA, INC.

PM – more often wrong than right…

SOLUCIA, INC.

“The year 1930, as a whole, should prove at least a fairly good year.”

-- *Harvard Economic Service, December 1929*

Why do it? Potential Use of Models

SOLUCIA, INC.

Medical Management Perspective

- Identifying individuals at very high risk of an event (death, LTC, disability, annuity surrender, etc.).
- Identify management opportunities and determine resource allocation / prioritization.

Identification – how?

SOLUCIA, INC.

- The art and science of predictive modeling!
- There are many different algorithms for identifying member conditions. THERE IS NO SINGLE AGREED FORMULA.
- Condition identification often requires careful balancing of sensitivity and specificity.

Identification – example (Diabetes)

SOLUCIA, INC.

Inpatient Hospital Claims – ICD-9 Claims Codes

ICD-9-CM CODE DIABETES	DESCRIPTION
250.xx	Diabetes mellitus
357.2	Polyneuropathy in diabetes
362.0, 362.0x	Diabetic retinopathy
366.41	Diabetic cataract
648.00-648.04	Diabetes mellitus (as other current condiition in mother classifiable elsewhere, but complicating pregnancy, childbirth or the puerperioum.

Diabetes – additional codes

SOLUCIA, INC.

CODES	CODE TYPE	DESCRIPTION - ADDITIONAL
DIABETES;		
G0108, G0109	HCPCS	Diabetic outpatient self-management training services, individual or group
J1815	HCPCS	Insulin injection, per 5 units
67227	CPT4	Destruction of extensive or progressive retinopathy, (e.g. diabetic retinopathy) one or more sessions, cryotherapy, diathermy
67228	CPT4	Destruction of extensiive or progressive retinopathy, one or more sessions, photocoagulation (laser or xenon arc).
996.57	ICD-9-CM	Mechanical complications, due to insulin pump
V45.85	ICD-9-CM	Insulin pump status
V53.91	ICD-9-CM	Fitting/adjustment of insulin pump, insulin pump titration
V65.46	ICD-9-CM	Encounter for insulin pump training

Diabetes – drug codes

SOLUCIA, INC.

Insulin or Oral Hypoglycemic Agents are often used to identify members. A simple example follows; for more detail, see the HEDIS code-set.

Insulin	
2710*	Insulin**

OralAntiDiabetics	
2720*	Sulfonylureas**
2723*	Antidiabetic - Amino Acid Derivatives**
2725*	Biguanides**
2728*	Meglitinide Analogues**
2730*	Diabetic Other**
2740*	ReductaseInhibitors**
2750*	Alpha-Glucosidase Inhibitors**
2760*	Insulin Sensitizing Agents**
2799*	Antiadiabetic Combinations**

All people are not equally identifiable

SOLUCIA, INC.

Definition Examples:

Narrow:	**Hospital Inpatient (primary Dx); Face-to-face professional (no X-Ray; Lab)**
Broad:	**Hospital I/P (any Dx); All professional**
Rx:	**Narrow + Outpatient Prescription**

Prevalence of 5 Chronic conditions

	Narrow	Broad	Rx
Medicare	24.4%	32.8%	30.8%
Commercial	4.7%	6.3%	6.6%

Identification: False Positives/ False Negatives

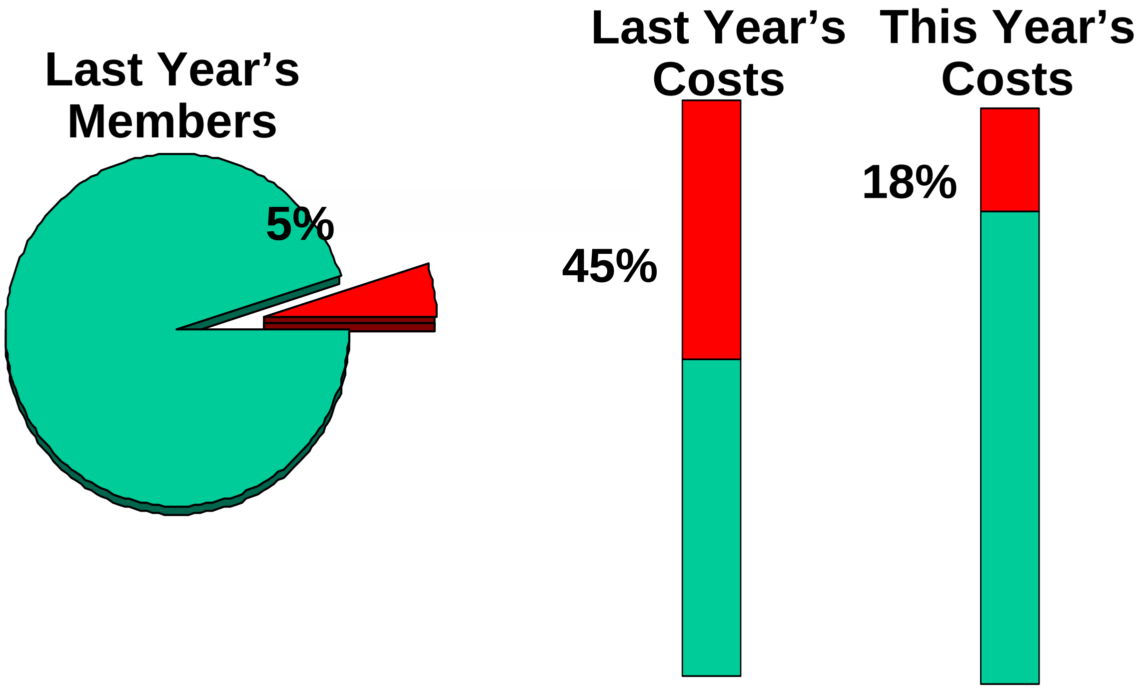
SOLUCIA, INC.

False Positive Identification Incidence through Claims
Medicare Advantage Population (with drug benefits)
Diabetes Example

Year 2 \ Year 1	Narrow	+ Broad	+ Rx	TOTAL
Narrow	75.9%			
+ Broad		85.5%		
+ Rx			92.6%	
Not Identified	24.1%	14.5%	7.4%	
TOTAL	100.0%	100.0%	100.0%	100.0%

Prospective versus Retrospective Targeting

SOLUCIA, INC.



Cost Stratification of a Large Population

	0.0% - 0.5%	0.5% - 1.0%	Top 1%	Top 5%	Total
Population	67,665	67,665	135,330	676,842	13,537,618
Actual Cost	$3,204,433,934	$1,419,803,787	$4,624,237,721	$9,680,579,981	$21,973,586,008
PMPY Total Actual Cost	$47,357	$20,977	$34,170	$14,303	$1,623
Percentage of Total Cost	**14.6%**	**6.5%**	**21.1%**	44.1%	100%
Patients with > $50,000 in Claims					
	0.0% - 0.5%	**0.5% - 1.0%**	**Top 1%**	**Top 5%**	**Total**
Number of Patients	19,370	5,249	24,619	32,496	35,150
Percentage of Total	55.1%	14.9%	**70.0%**	92.4%	100.0%

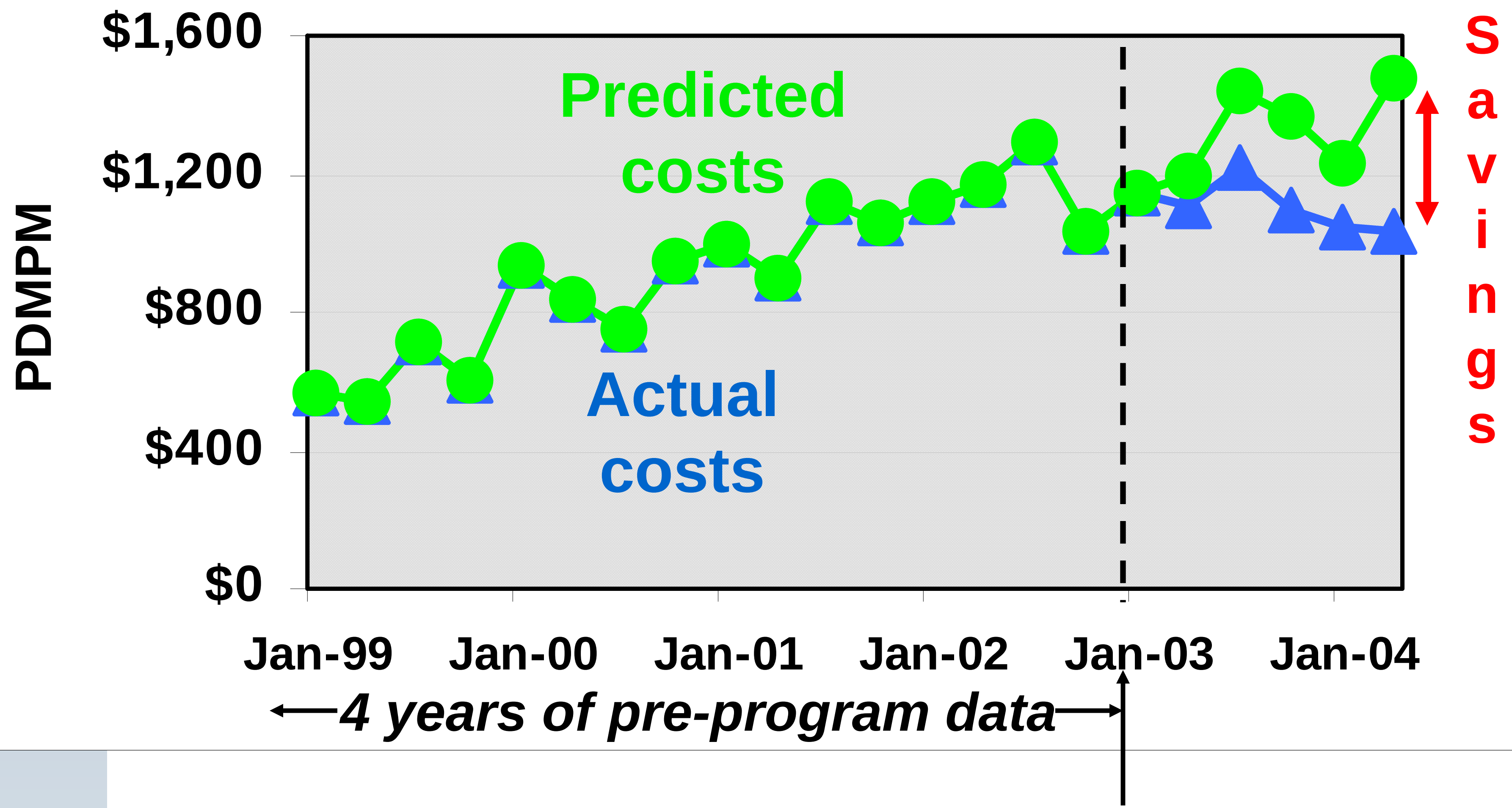
Why do it? Potential Use of Models

SOLUCIA, INC.

Program Evaluation / Reimbursement Perspective

- Predicting *what would have happened* absent a program.
- Predicting resource use in the "typical" population.

Example 1: Time Series



Example 2: Normalized resources

SOLUCIA, INC.

Member ID	Single Condition	RiskScoreID	PgmCode	NonDup Patient Count	Patient Count x Risk Score	Expected Claims Cost
1080	CHF	39.8	200	1	39.774	$ 58,719
532	Cancer 1	174.2	100	1	174.189	210,829
796	Cancer 2 + Chronic cond.	159.7	100	1	159.671	1,289,469
531	Cancer 2 + No Chron. cond	135.3	100	1	135.289	338,621
1221	Multiple Chron conds.	28.8	200	1	28.811	34,660
710	Acute conds and Chron	110.9	100	1	110.87	100,547
795	Acute conds and Chron	121.1	100	1	121.083	148,107
882	Diabetes	25.7	200	1	25.684	22,647
967	Cardiac	24.5	200	1	24.465	1,308
881	Asthma	24.1	200	1	24.096	15,776
						$ 2,220,683

Why do it? Potential Uses of Models

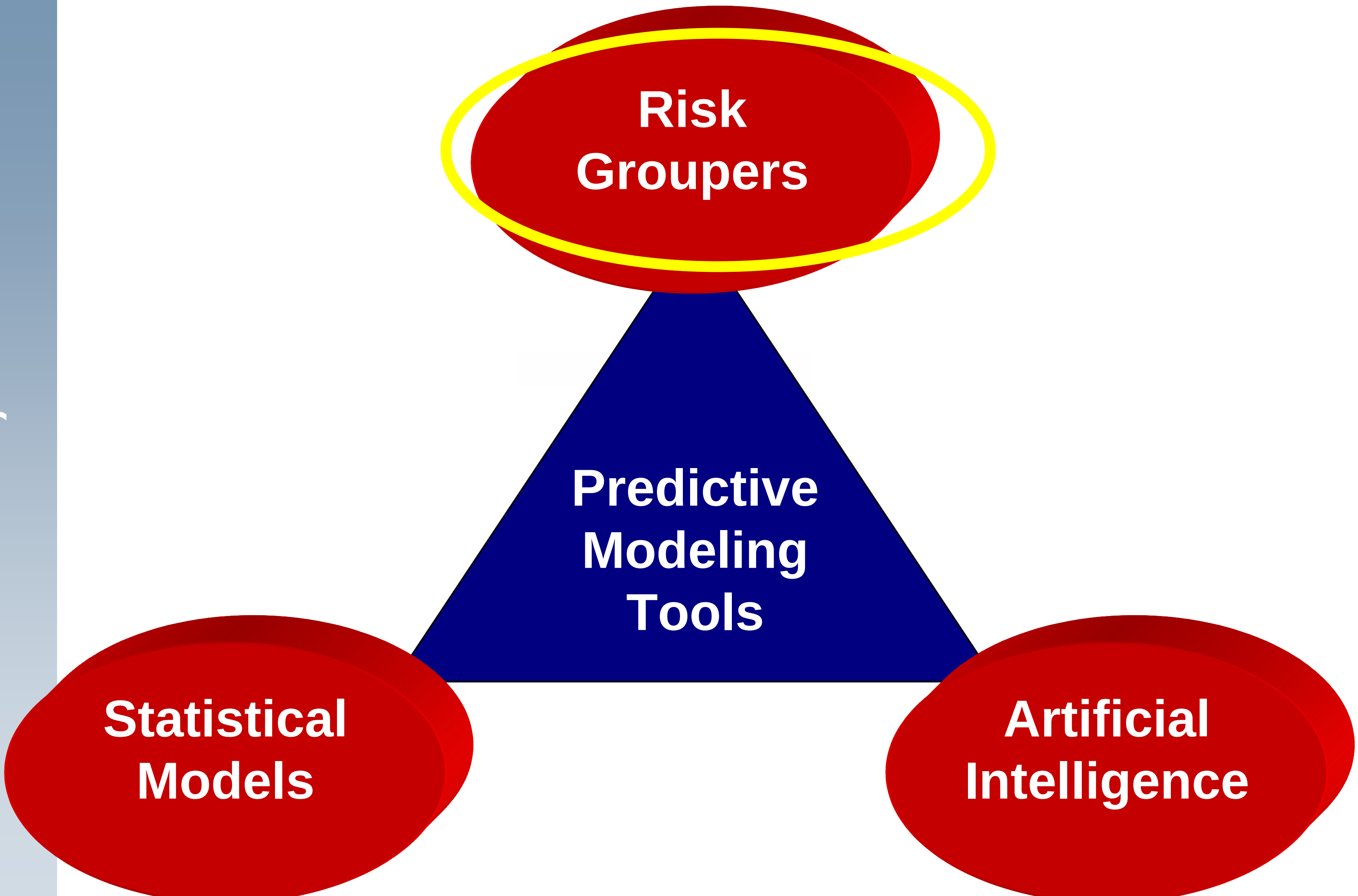
SOLUCIA, INC.

Actuarial, Underwriting and Profiling Perspectives

- Calculating renewal premium
- Profiling of provider
- Provider & health plan contracting

Types of Predictive Modeling Tools

SOLUCIA, INC.



Uses of Risk Groupers

Risk Groupers can be used for these 3 purposes ... but best for actuarial, underwriting and profiling

- Actuarial, Underwriting and Profiling Perspectives
- Medical Management Perspective
- Program Evaluation Perspective

What are the different types of risk groupers?

Selected Risk Groupers

Company	Risk Grouper	Data Source
IHCIS/Ingenix	ERG	Age/Gender, ICD-9 NDC, Lab
UC San Diego	CDPS	Age/Gender, ICD -9 NDC
DxCG	DCG RxGroup	Age/Gender, ICD -9 Age/Gender, NDC
Symmetry/Ingenix	ERG PRG	ICD – 9, NDC NDC
Johns Hopkins	ACG	Age/Gender, ICD – 9

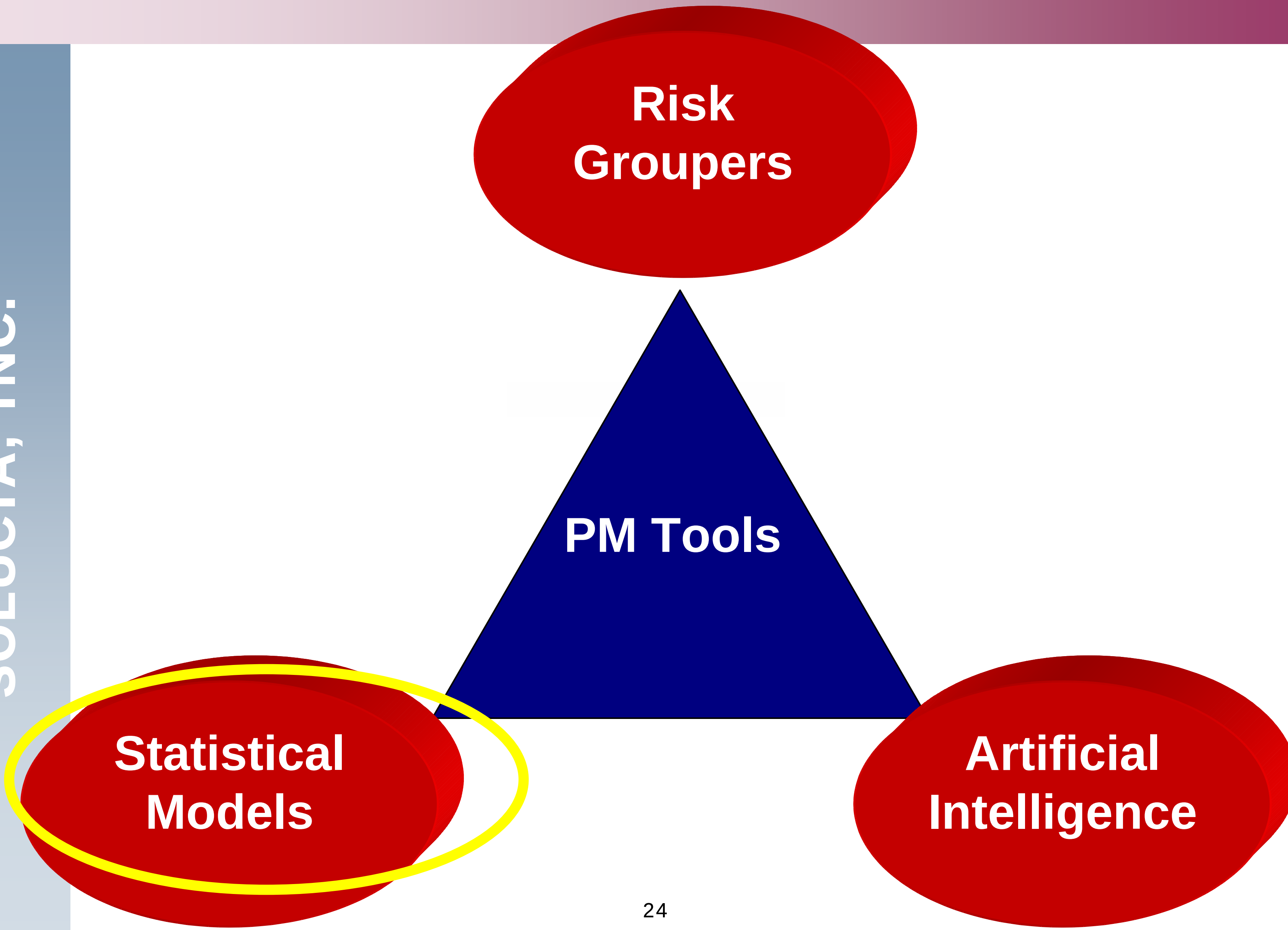
Risk Grouper Summary

SOLUCIA, INC.

1. Similar performance among all leading risk groupers*.
2. Risk grouper modeling tools use *different algorithms* to group the source data.
3. Risk groupers use *relatively limited data* sources (e.g. DCG and Rx Group use ICD-9 and NDC codes but not lab results or HRA information)
4. Most Risk Grouper based Predictive Models combine also use statistical analysis.

*** See New SOA study (Winkelman et al) published this year. Available from SOA.**

Types of Predictive Modeling Tools



Uses of Statistical Models

Statistical models can be used for all 3 uses

Medical Management Perspective	Actuarial, Underwriting and Profiling Perspectives	Program Evaluation Perspective

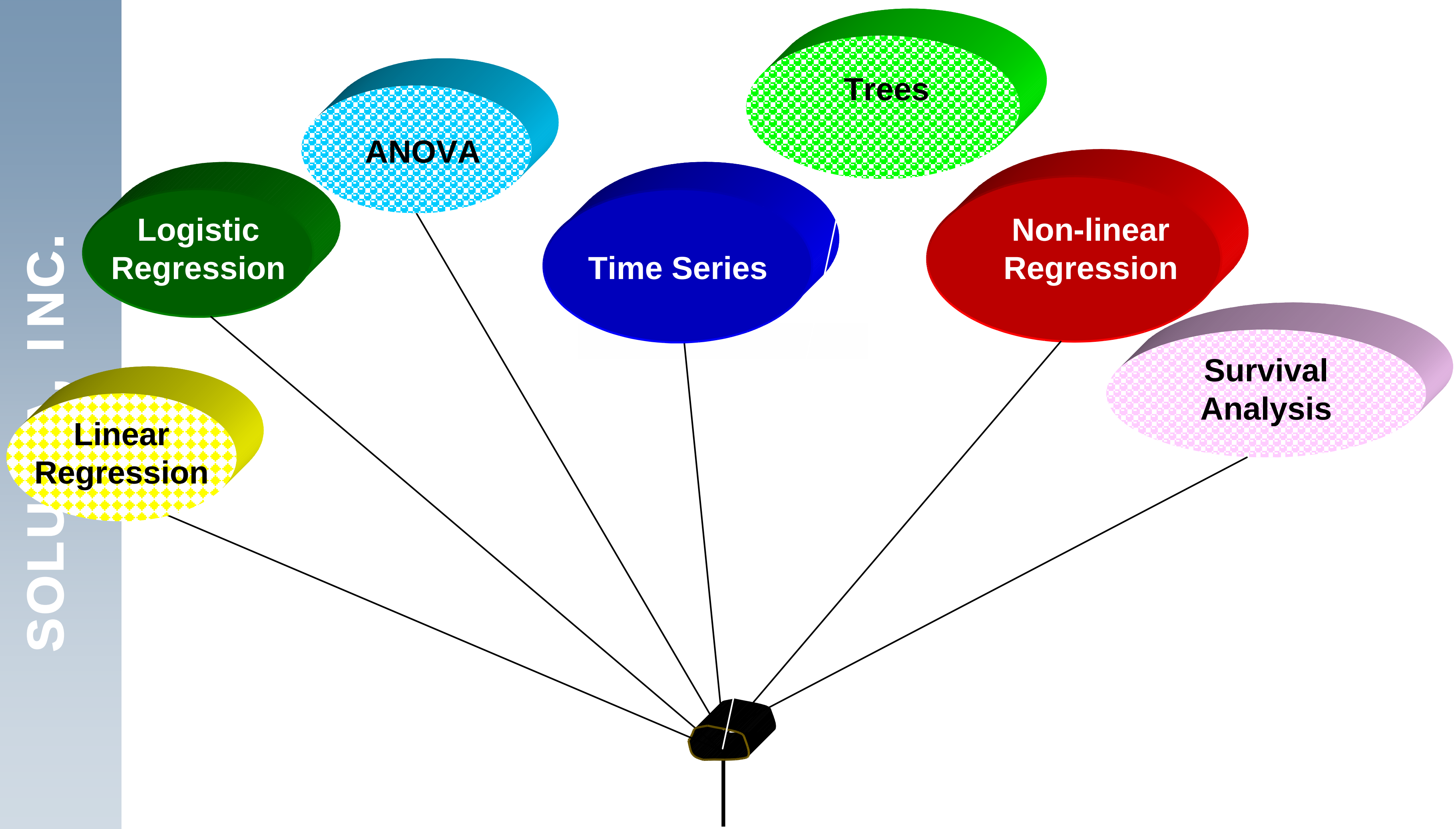
SOLUC

Statistical Models

SOLUCIA, INC.

What are the different types of statistical models?

Types of Statistical Models



Multiple Regression Model Example

Finding	Hierarchy	Coefficient	Notes
Diabetes	Low cost DM	0	Trumped by Hi cost
Diabetic nephropathy	Hi cost DM	2.455	
Angina	Low cost CAD	0	Trumped by Hi cost
Migraines	Med cost headache	0.208	
	Subtotal	2.763	
Age-related base		0.306	
Gender-related base		-0.087	
	Risk	2.982	

Statistical Models

SOLUCIA, INC.

Time series modeling tools is another type of statistical modeling tool – it requires a lot of historical data.

Time Series

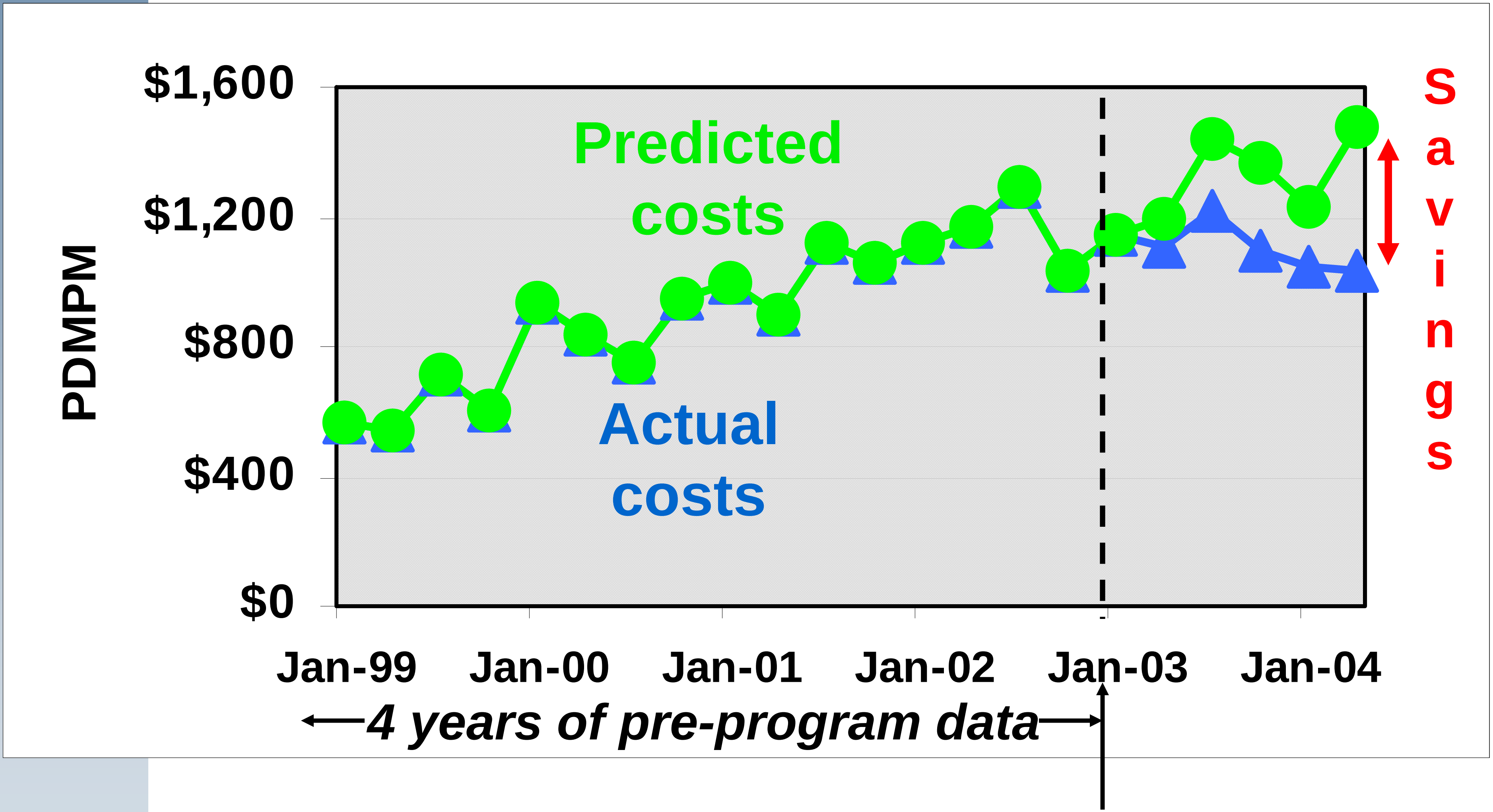
SOLUCIA, INC.

Time series analysis is to

a) Identify the pattern of observed time series data and

b) Forecast future values by extrapolating the identified pattern.

Example: Time Series

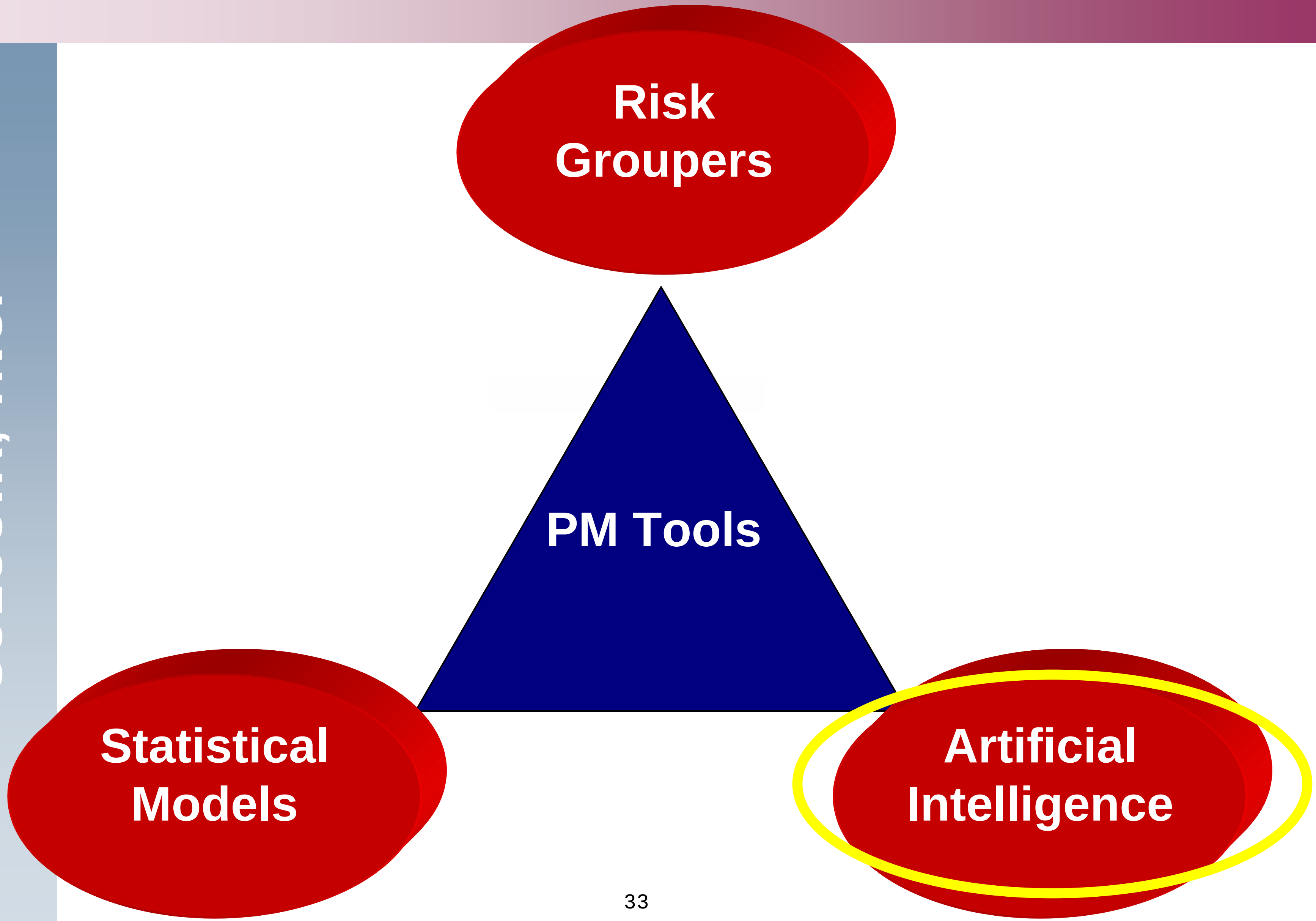


Statistical Model Summary

SOLUCIA, INC.

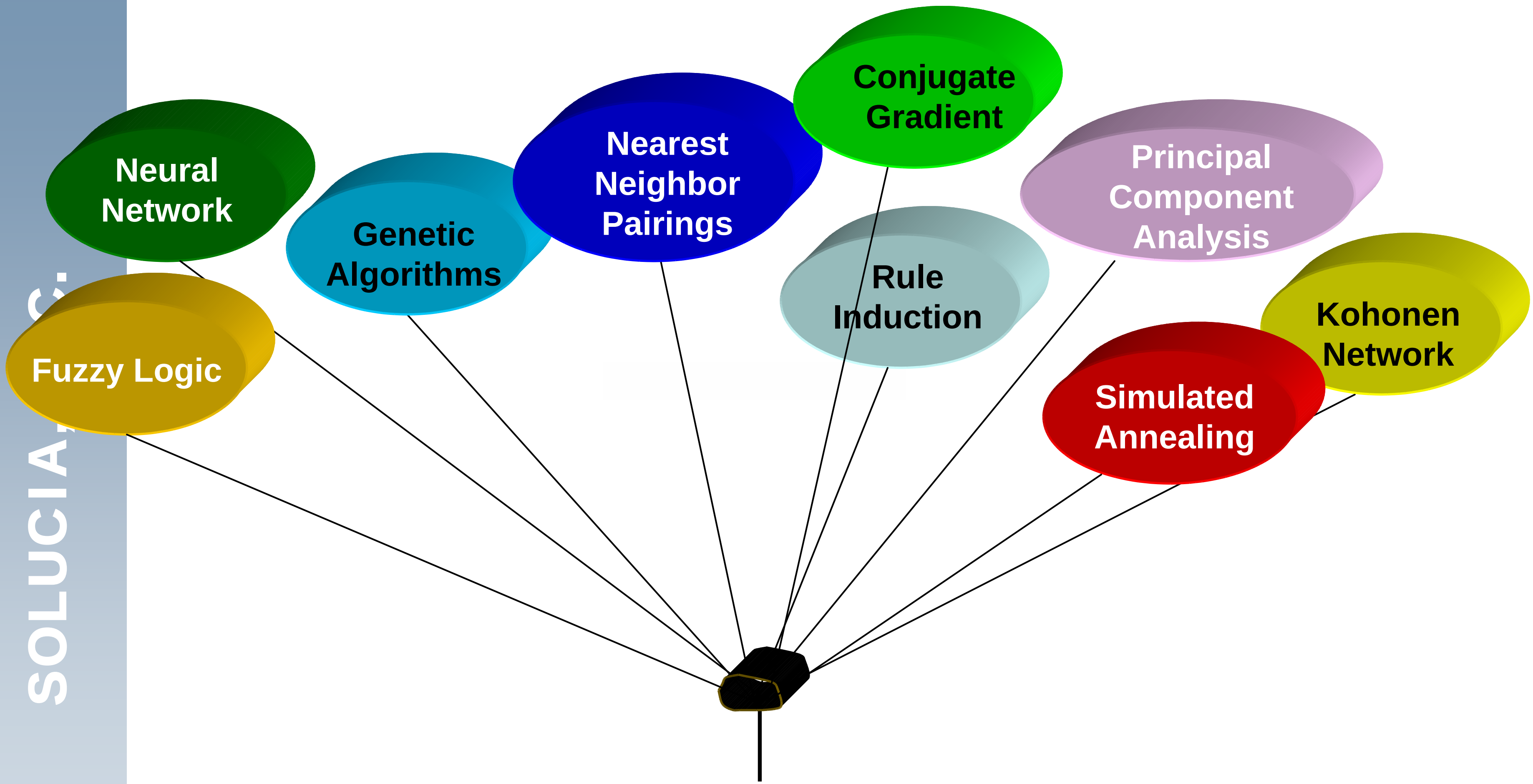
1. Statistical models can be used for a number of actuarial applications: evaluation, premium calculation, provider profiling and resource allocation.
2. The predictive model is a critical component of successful medical management intervention programs - "impactability is key in medical management".
3. Statistical models can use all available detailed data (e.g. lab results or HRA).

Types of Predictive Modeling Tools



What are the different types of artificial intelligence models?

Artificial Intelligence Models



Features of Neural Networks

SOLUCIA, INC.

Reality

NN tracks complex relationships by resembling the human brain

Perception

NN can accurately model complicated health care systems

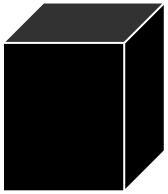
Reality

- Performance equals standard statistical models
- Models overfit data

Neural Network Summary

SOLUCIA, INC.

1. Good academic approach.
2. Few data limitations.
3. Performance comparable to other approaches.
4. Can be hard to understand the output of neural networks (black box).



In Summary

SOLUCIA, INC.

1. Leading predictive modeling tools have similar performance.
2. Selecting a predictive modeling tool should be based on your specific objectives - one size doesn't fit all.
3. A good predictive model for medical management should be linked to the intervention (e.g. impactability).
4. "Mixed" models can increase the power of a single model.

PM is <u>NOT</u> always about *Cost Prediction…..*

SOLUCIA, INC.

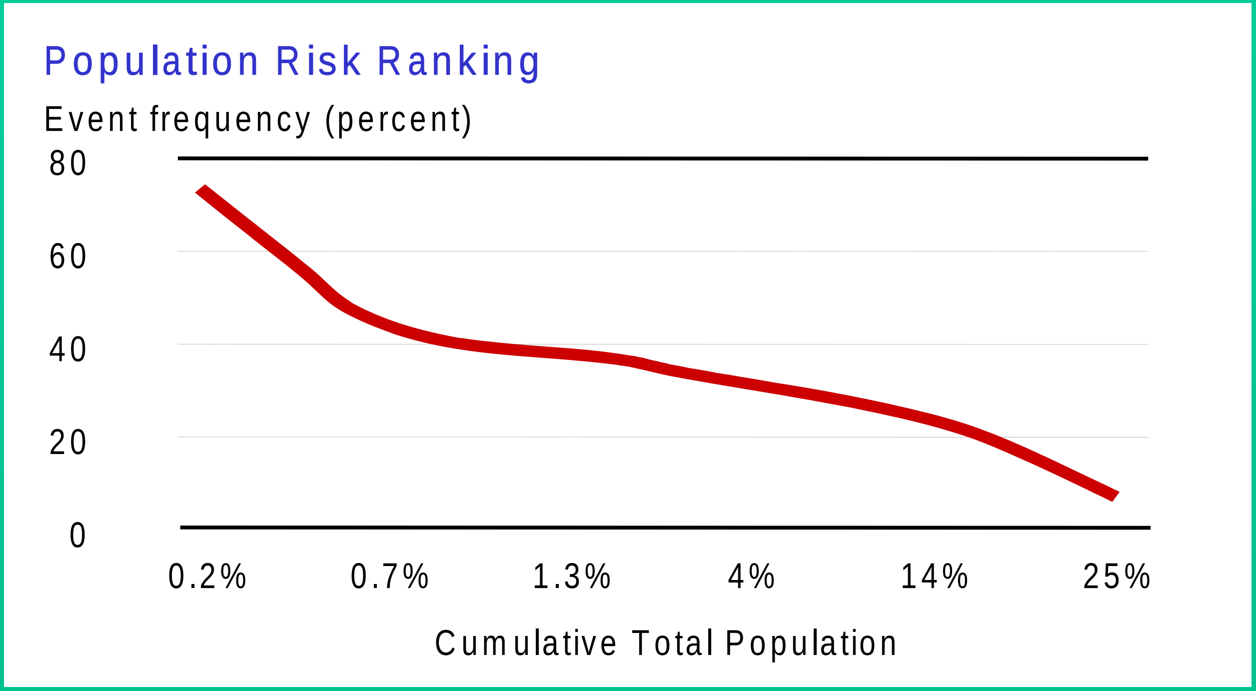
…..it IS about resource allocation.

- **Where/how should you allocate resources?**
- **Who is *intervenable* or *impactable*?**
- **What can you expect for outcomes?**
- **How can you manage the key drivers of the economic model for better outcomes?**

Remember this chart?

	0.0% - 0.5%	0.5% - 1.0%	Top 1%	Top 5%	Total
Population	67,665	67,665	135,330	676,842	13,537,618
Actual Cost	$3,204,433,934	$1,419,803,787	$4,624,237,721	$9,680,579,981	$21,973,586,008
PMPY Total Actual Cost	$47,357	$20,977	$34,170	$14,303	$1,623
Percentage of Total Cost	**14.6%**	**6.5%**	**21.1%**	44.1%	100%
Patients with > $50,000 in Claims					
	0.0% - 0.5%	**0.5% - 1.0%**	**Top 1%**	**Top 5%**	**Total**
Number of Patients	19,370	5,249	24,619	32,496	35,150
Percentage of Total	55.1%	14.9%	**70.0%**	92.4%	100.0%

Decreasing Cost / Decreasing Opportunity



SOLUCIA, INC.

Economic Model: Simple example

SOLUCIA, INC.

- 30,000 eligible members (ee/dep)
- 1,500 – 2,000 with chronic conditions
- 20% "high risk" – 300 to 400
- 60% are reachable and enroll: 180 - 240
- Admissions/high-risk member/year: 0.65
- "Change behavior" of 25% of these:
- - reduced admissions: 29 to 39 annually
- - cost: $8,000/admission
- Gross Savings: $232,000 to $312,000
- - $0.64 to $0.87 pmpm.

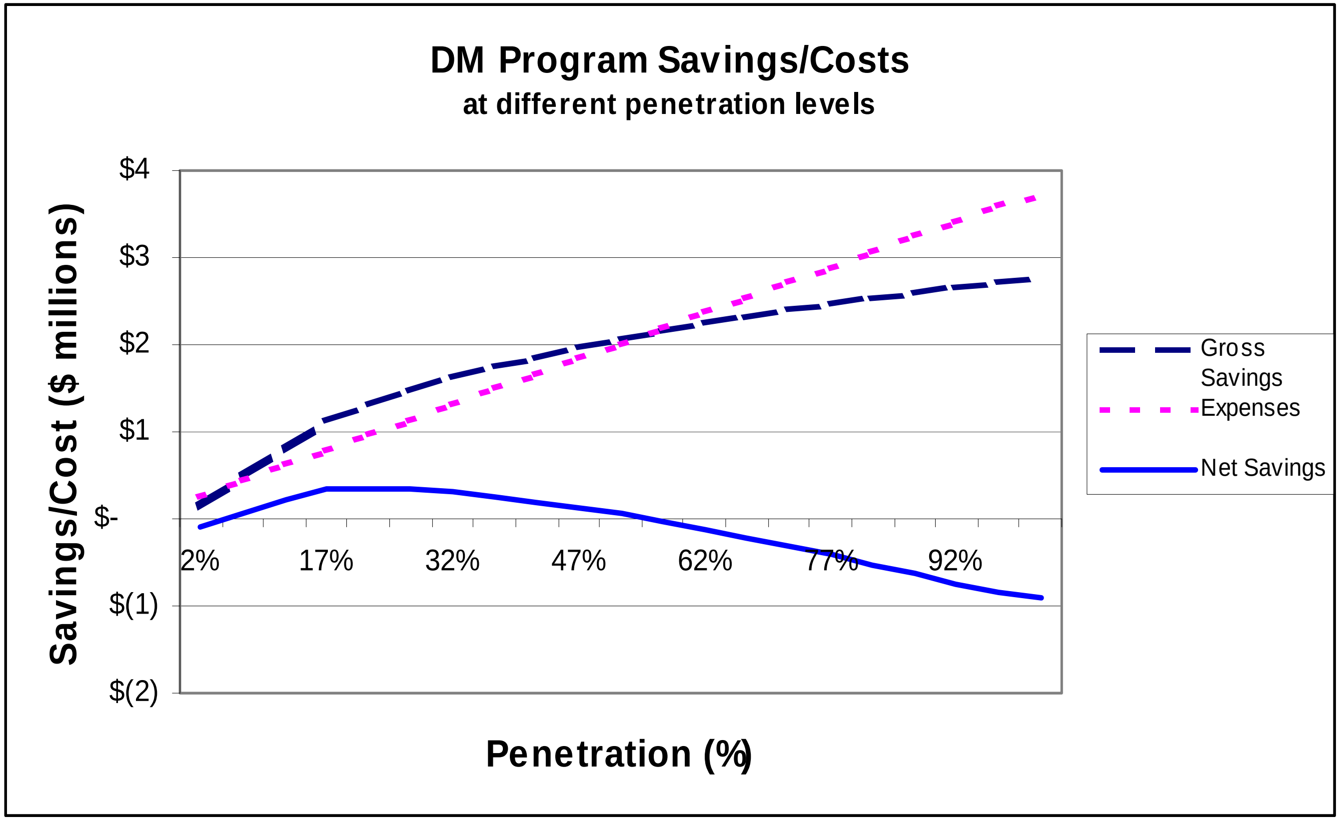
Key drivers of the economic model

SOLUCIA, INC.

- Prevalence within the population (numbers)
- Ability to Risk Rank the Population
- Data quality
- Reach/engage ability
- Cost/benefit of interventions
- Timeliness
- Resource productivity
- Random variability in outcomes

Understanding the Economics

SOLUCIA, INC.



SOLUCIA, INC.

Modeling

What is a model?

- A model is a set of coefficients to be applied to production data in a live environment.

- With individual data, the result is often a predicted value or “score”. For example, the likelihood that an individual will purchase something, or will experience a high-risk event (surrender; claim, etc.).

- For underwriting, we can predict either cost or risk-score.

Practical Example of Model-Building

SOLUCIA, INC.

Background

Available data for creating the score included the following

- Eligibility/demographics
- Rx claims
- Medical claims

For this project, several data mining techniques were considered: neural net, CHAID decision tree, and regression. The regression was chosen for the following reasons:

With proper data selection and transformation, the regression was very effective, more so than the tree.

1. Split the dataset randomly into halves

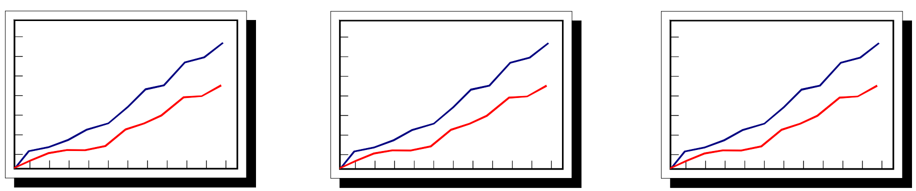
SOLUCIA, INC.

Master Dataset

Analysis Dataset

Test Dataset

Diagnostics



Put
scor
data
are

2. Build and Transform independent variables

- In any data-mining project, the output is only as good as the input.
- Most of the time and resources in a data mining project are actually used for variable preparation and evaluation, rather than generation of the actual “recipe”.

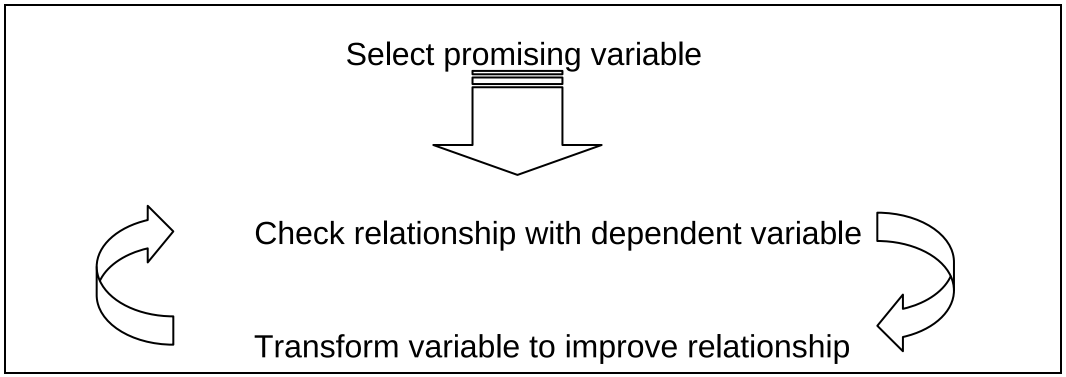
3. Build composite dependent variable

SOLUCIA, INC.

- A key step is the choice of dependent variable. What is the best choice?
- A likely candidate is total patient cost in the predictive period. But total cost has disadvantages
  - It includes costs such as injury or maternity that are not generally predictable.
  - It includes costs that are steady and predictable, independent of health status (capitated expenses).
  - It may be affected by plan design or contracts.
- We generally predict total cost (allowed charges) net of random costs and capitated expenses.
- Predicted cost can be converted to a risk-factor.

3. Build and transform Independent Variables

The process below is applied to variables from the baseline data.

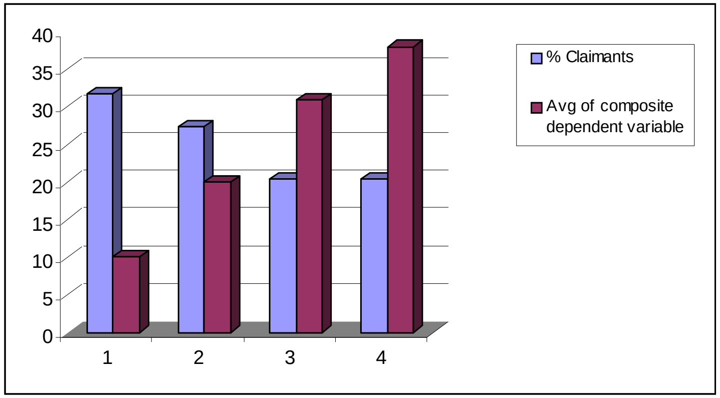


- Typical transforms include
- Truncating data ranges to minimized the effects of outliers.
- Converting values into binary flag variables.
- Altering the shape of the distribution with a log transform to compare orders of magnitude.
- Smoothing progression of independent variables

3. Build and transform Independent Variables

SOLUCIA, INC.

- A simple way to look at variables
- Convert to a discrete variable. Some variables such as number of prescriptions are already discrete. Real-valued variables, such as cost variables, can be grouped into ranges
- Each value or range should have a significant portion of the patients.
- Values or ranges should have an ascending or descending relationship with average value of the composite dependent variable.



Typical "transformed" variable

4. Select Independent Variables

SOLUCIA, INC.

- The following variables were most promising
- Age -Truncated at 15 and 80
- Baseline cost
- Number of comorbid condition truncated at 5
- MClass
  - Medical claims-only generalization of the comorbidity variable.
  - Composite variable that counts the number of distinct ICD9 ranges for which the claimant has medical claims.
  - Ranges are defined to separate general disease/condition categories.
- Number of prescriptions truncated at 10

4. Select Independent Variables (contd.)

SOLUCIA, INC.

- Scheduled drug prescriptions truncated at 5
- NClass
  - Rx-only generalization of the comorbidity variable.
  - Composite variable that counts the number of distinct categories distinct ICD9 ranges for which the claimant has claims.
  - Ranges are defined using GPI codes to separate general disease/condition categories.
- Ace inhibitor flag
- Neuroleptic drug flag
- Anticoagulants flag
- Digoxin flag
- Diuretics flag
- Number of corticosteroid drug prescriptions truncated at 2

5. Run Stepwise Linear Regression

SOLUCIA, INC.

An ordinary linear regression is simply a formula for determining a best-possible linear equation describing a dependent variable as a function of the independent variables. But this pre-supposes the selection of a best-possible set of independent variables. How is this best-possible set of independent variables chosen?

One method is a stepwise regression. This is an algorithm that determines both a set of variables and a regression. Variables are selected in order according to their contribution to incremental R^2

5. Run Stepwise Linear Regression (continued)

SOLUCIA, INC.

Stepwise Algorithm

1. Run a single-variable regression for each independent variable. Select the variable that results in the greatest value of R^2. This is "Variable 1".

2. Run a two-variable regression for each remaining independent variable. In each regression, the other independent variable is Variable 1. Select the remaining variable that results in the greatest <u>incremental</u> value of R^2. This is "Variable 2."

3. Run a three-variable regression for each remaining independent variable. In each regression, the other two independent variables are Variables 1 and 2. Select the remaining variable that results in the greatest <u>incremental</u> value of R^2. This is "Variable 3."

……

n. Stop the process when the incremental value of R^2 is below some pre-defined threshold.

6. Results - Examples

- Stepwise linear regressions were run using the "promising" independent variables as inputs and the composite dependent variable as an output.
- Separate regressions were run for each patient sex.
- Sample Regressions

  - Female
    - Scheduled drug prescription 358.1
    - NClass 414.5
    - MClass 157.5
    - Baseline cost 0.5
    - Diabetes Dx 1818.9
    - Intercept 18.5

Why are some variables selected while others are omitted? The stepwise algorithm favors variables that are relatively uncorrelated with previously-selected variables. The variables in the selections here are all relatively independent of each other.

6. Results - Examples

- Examples of application of the female model

Female Regression Regression Formula

(Scheduled Drug *358.1) + (NClass*414.5) + (Cost*0.5) + (Diabetes*1818.9) + (MClass*157.5) -18.5

Claimant ID	Raw Value	Transformed Value	Predicted Value	Actual Value
Schedule Drugs				
1	3	2	$ 716.20	
2	2	2	$ 716.20	
3	0	1	$ 358.10	
NClass				
1	3	3	$ 1,243.50	
2	6	6	$ 2,487.00	
3	0	0.5	$ 207.25	
Cost				
1	423	2,000	$ 1,000.00	
2	5,244	6,000	$ 3,000.00	
3	1,854	2,000	$ 1,000.00	
Diabetes				
1	0	0	$ -	
2	0	0	$ -	
3	0	0	$ -	
MClass				
1	8	3	$ 472.50	
2	3	2	$ 315.00	
3	0	0.5	$ 78.75	
TOTAL				
1			$ 3,413.70	$ 4,026.00
2			$ 6,499.70	$ 5,243.00
3			$ 1,625.60	$ 1,053.00

Transform Function

	Schedule Drugs		
Value Range	RV< 2	2 < RV < 5	RV >5
Transformed Value	1.0	2.0	3.0

	NClass		
Value Range	RV < 2	2 < RV < 5	RV > 5
Transformed Value	0.5	3.0	6.0

	Cost		
Value Range	RV < 5k	5k < RV < 10k	RV > 10k
Transformed Value	2,000	6,000	10,000

	Diabetes	
Value Range	Yes	No
Transformed Value	1.0	0.0

	MClass		
Value Range	RV < 1	1 < RV < 7	RV > 7
Transformed Value	0.5	2.0	3.0

SOLUCIA, INC.

SOLUCIA, INC.

Model Modifications

Expanding and Changing the Model

Expanding definitions

Models for separate populations

Models for varying renewal years

Form of output

Trend

SOLUCIA, INC.

Evaluation

EVALUATION - Testing

SOLUCIA, INC.

- Various statistics available for evaluation:
  - R-squared
  - Mean Absolute Prediction Error
  - (Prediction – Actual) / Prediction
- Compare to existing tools
- Evaluate results and issues

Selected references

SOLUCIA, INC.

This is not an exhaustive bibliography. It is only a starting point for explorations.

- Shapiro, A.F. and Jain, L.C. (editors); *Intelligent and Other Computational Techniques in Insurance;* World Scientific Publishing Company; 2003.
- Dove, Henry G., Duncan, Ian, and Robb, Arthur; *A Prediction Model for Targeting Low-Cost, High-Risk Members of Managed Care Organizations*; The American Journal of Managed Care, Vol 9 No 5, 2003
- Berry, Michael J. A. and Linoff, Gordon; *Data Mining Techniques for Marketing, Sales and Customer Support*; John Wiley and Sons, Inc; 2004
- Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G Geoffrey; *Introduction to Linear Regression Analysis*; John Wiley and Sons, Inc; 2001
- Kahneman, Daniel, Slovic, Paul, and Tversky (editors); *Judgment under uncertainty: Heuristics and Biases;* Cambridge University Press; 1982

Selected references (contd.)

SOLUCIA, INC.

- Dove, Henry G., Duncan, Ian, and others; *Evaluating the Results of Care Management Interventions: Comparative Analysis of Different Outcomes Measures.* The SOA study of DM evaluation, available on the web-site at

  http://www.soa.org/professional-interests/health/hlth-evaluating-the-results-of-care-management-interventions-comparative-analysis-of-different-outcomes-measures-claims.aspx
- Winkelman R. and S. Ahmed. *A comparative analysis of Claims Based Methods of health risk assessment ofr Commercial Populations*. (2007 update to the SOA Risk-Adjuster study.) Available from the SOA; the 2002 study is on the website at: http://www.soa.org/files/pdf/_asset_id=2583046.pdf.

Further Questions?

SOLUCIA, INC.

iduncan@soluciaconsulting.com

Solucia Inc.
220 Farmington Avenue
Farmington, CT 06032

860-676-8808

www.soluciaconsulting.com