Prediction Markets: Do They Improve Risk Management?

Jon Schuler

MITRE Corporation

jschuler@mitre.org

703-983-3884

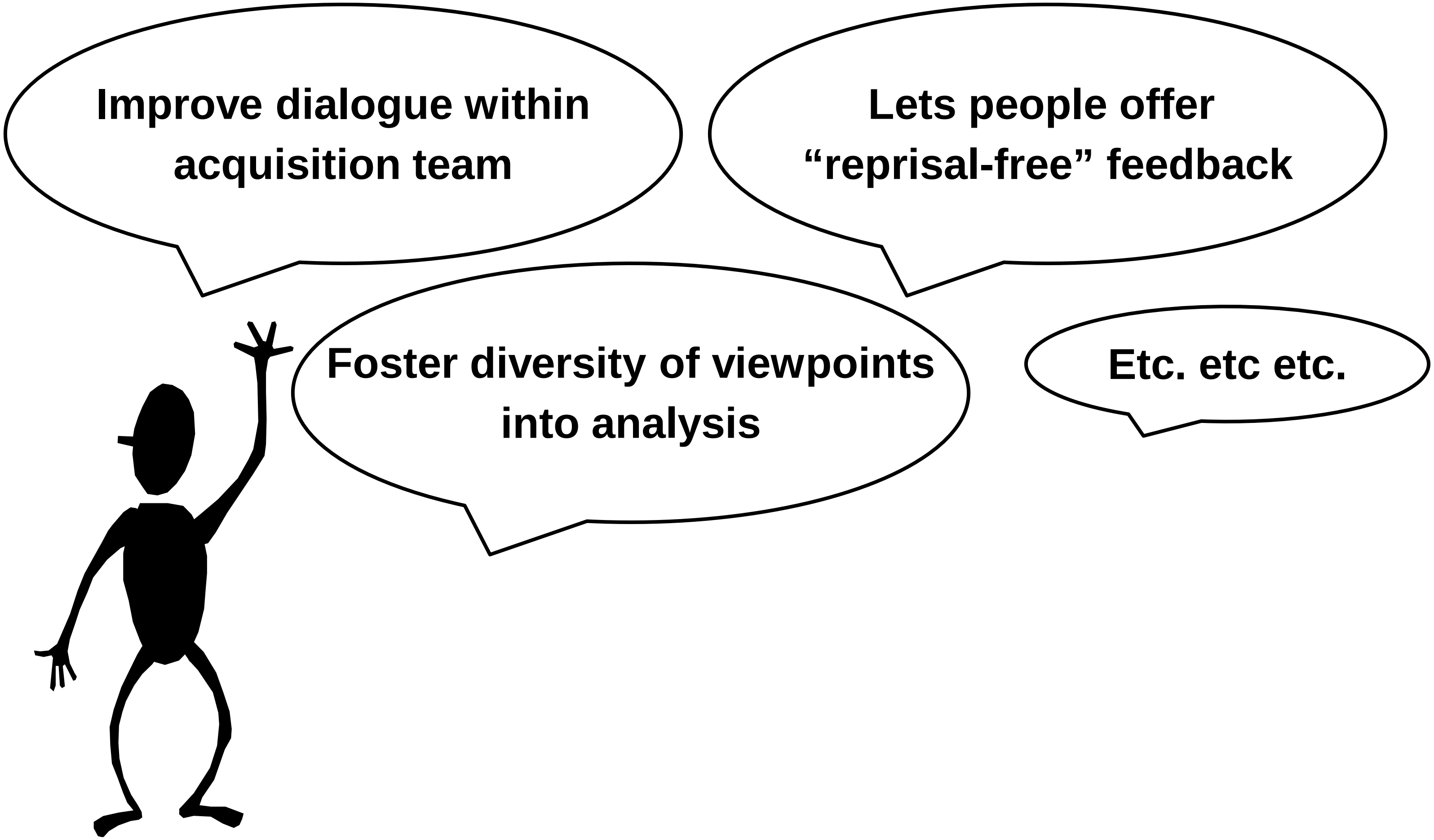
22 September 2010

MITRE

What we did:

- Introduced an *Inkling*™ prediction market to the USAF
- Generic questions:
  - "Will the New England Patriots sign Terrell Owens by the start of the 2010 NFL season? "
  - "Will both BP and U.S. Government sources officially report no more oil is leaking from the Deepwater Horizon well by July 1, 2010?
- Acquisition-program specific questions:
  - "Will Predator platforms adopt the Program2 Rev B specs?"
  - "Will any of the load-balancing, multi-path TRILL boxes tested at last month's UNH Interop lab appear in an USAF Advanced Technology Demonstration by Spring 2011?"
  - "Which of following Spectrum bands will Program2 be approved for use by Jan 2011"

“What’s a prediction market good for?”



Improve dialogue within acquisition team

Lets people offer “reprisal-free” feedback

Foster diversity of viewpoints into analysis

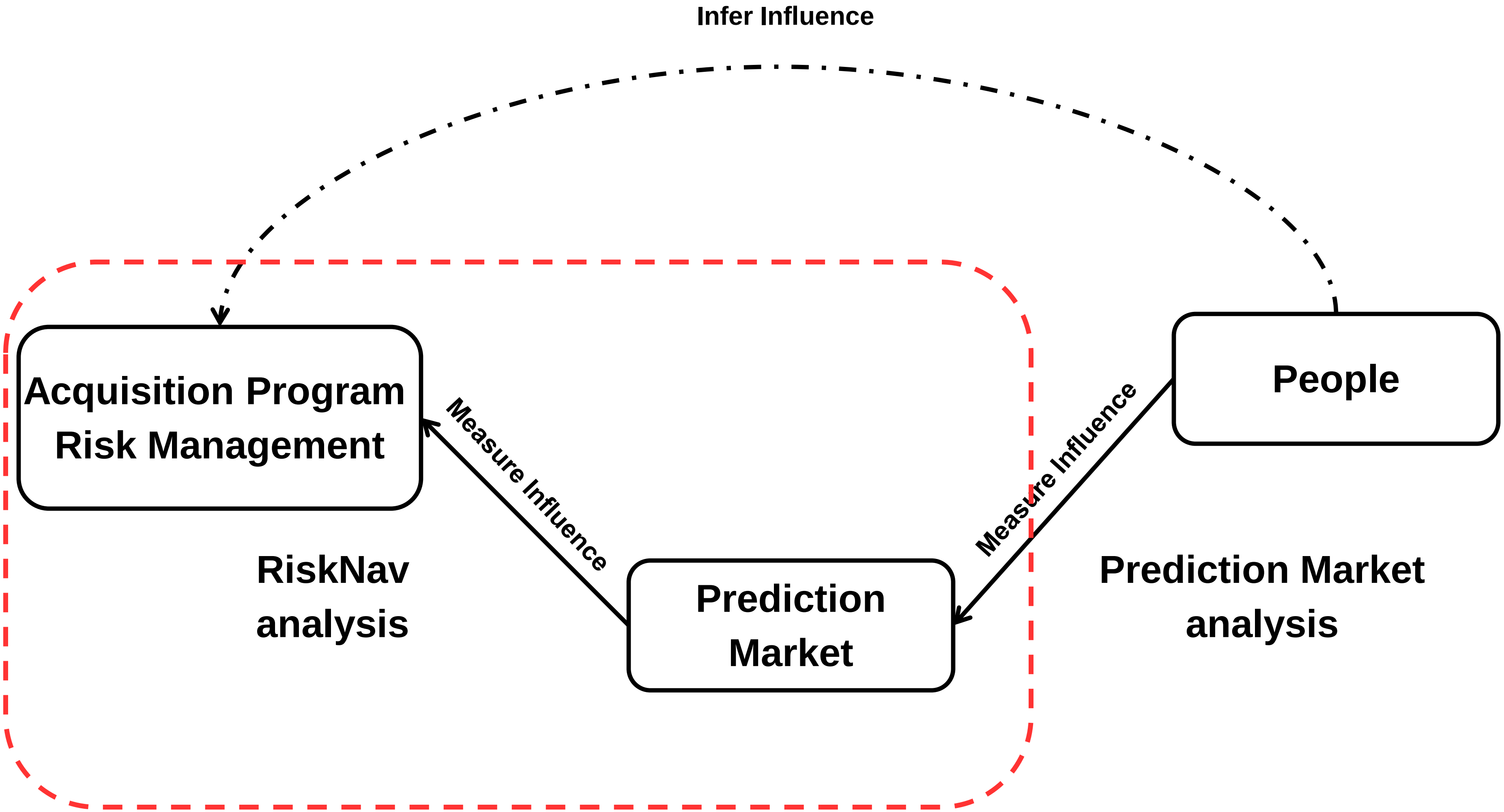
Etc. etc etc.

Can we clinically measure that?

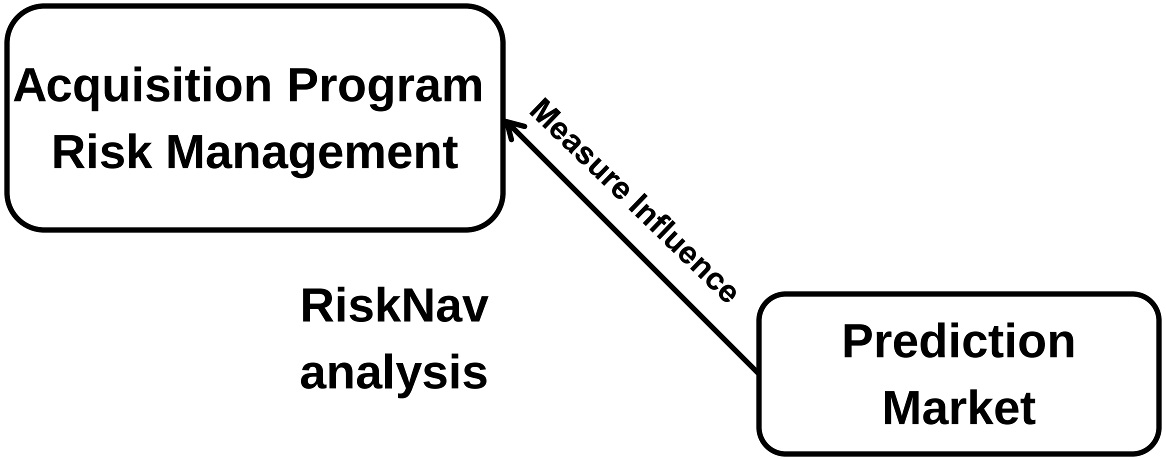
- Pre-existing risk-management process already in place
  - MITRE provides on-site support to the USAF
  - Established process that identifies, enunciates, quantifies, and mitigates acquisition risks
  - 14 acquisition programs; various levels of activity

- 2+ years data logged in RiskNav software
  - Front-end: web-based interface
    - Enter/modify assessments. Provide summary display
  - Back-end: Microsoft Access database
  - Primary use: characterize current state
  - Mining DB change-logs can extract a historical record

Prediction Market Analysis Framework

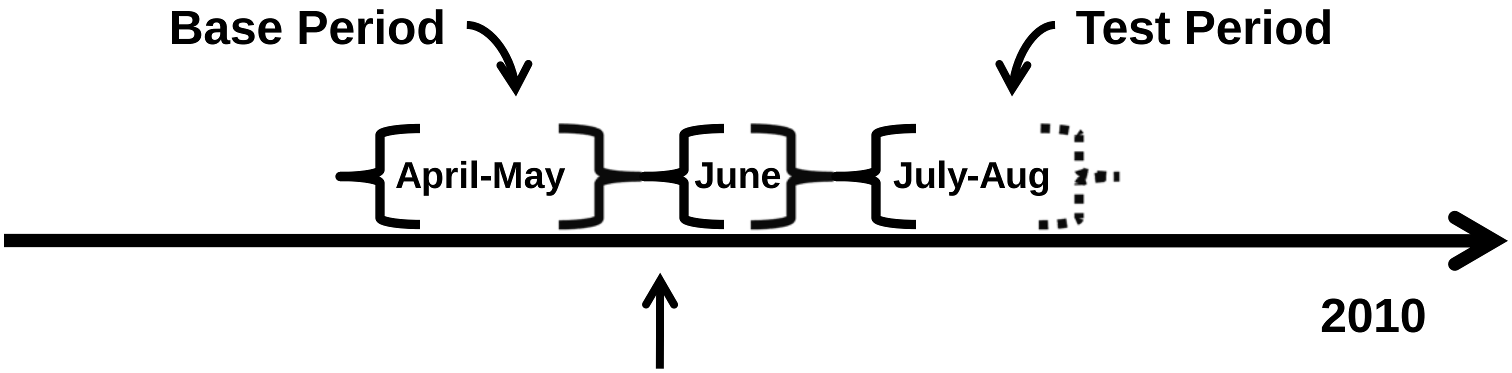


‘Treatment Effects’ we would expect to see



- Is there an increase in the overall database activity?
- Is there an increase in the rate of newly identified risks?
- Are new risks identified earlier from their event-horizon?
- Do risks get mitigated or closed more quickly?

Event timeline under analysis



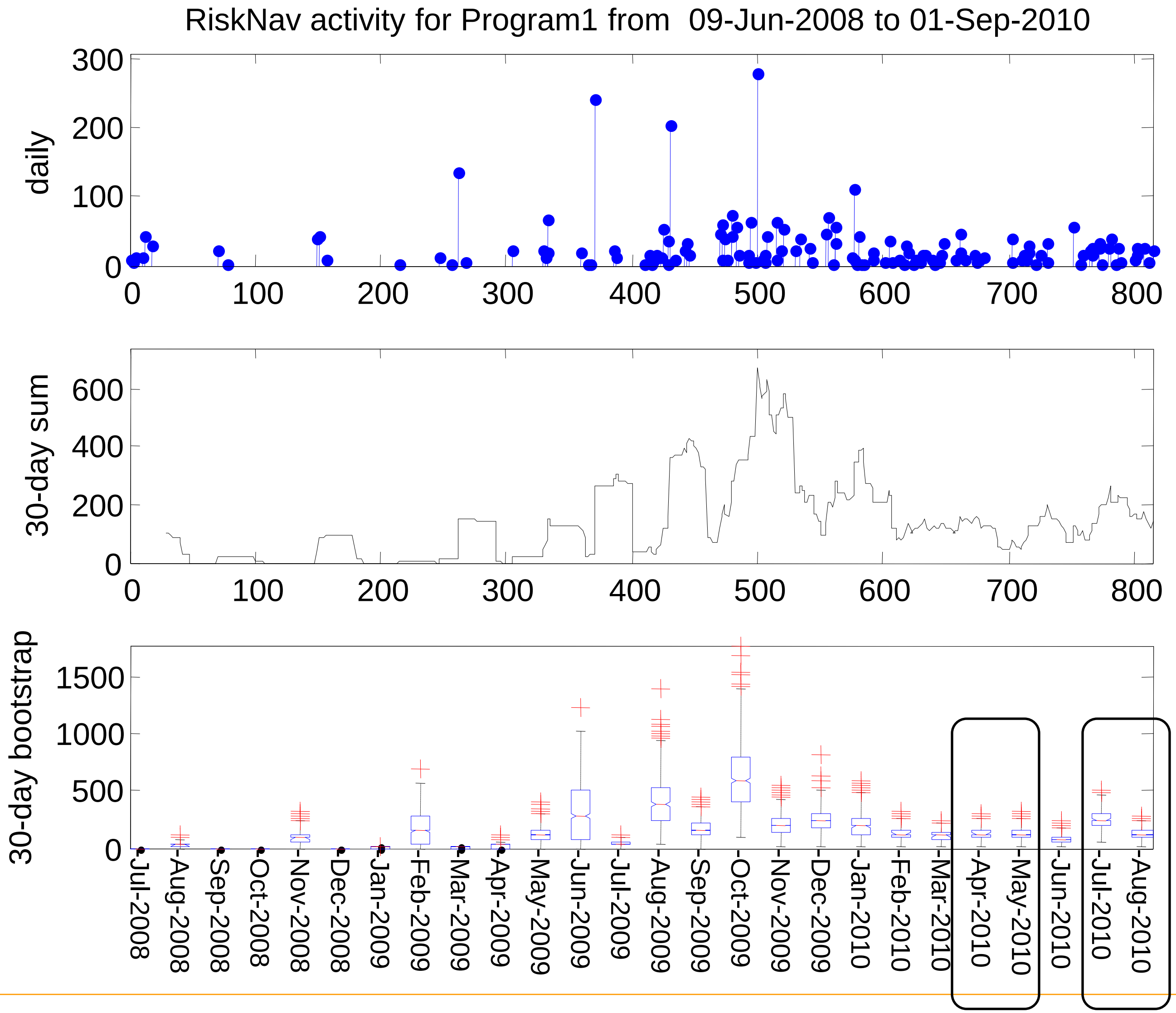
Inkling outreach/invitations

to USAF

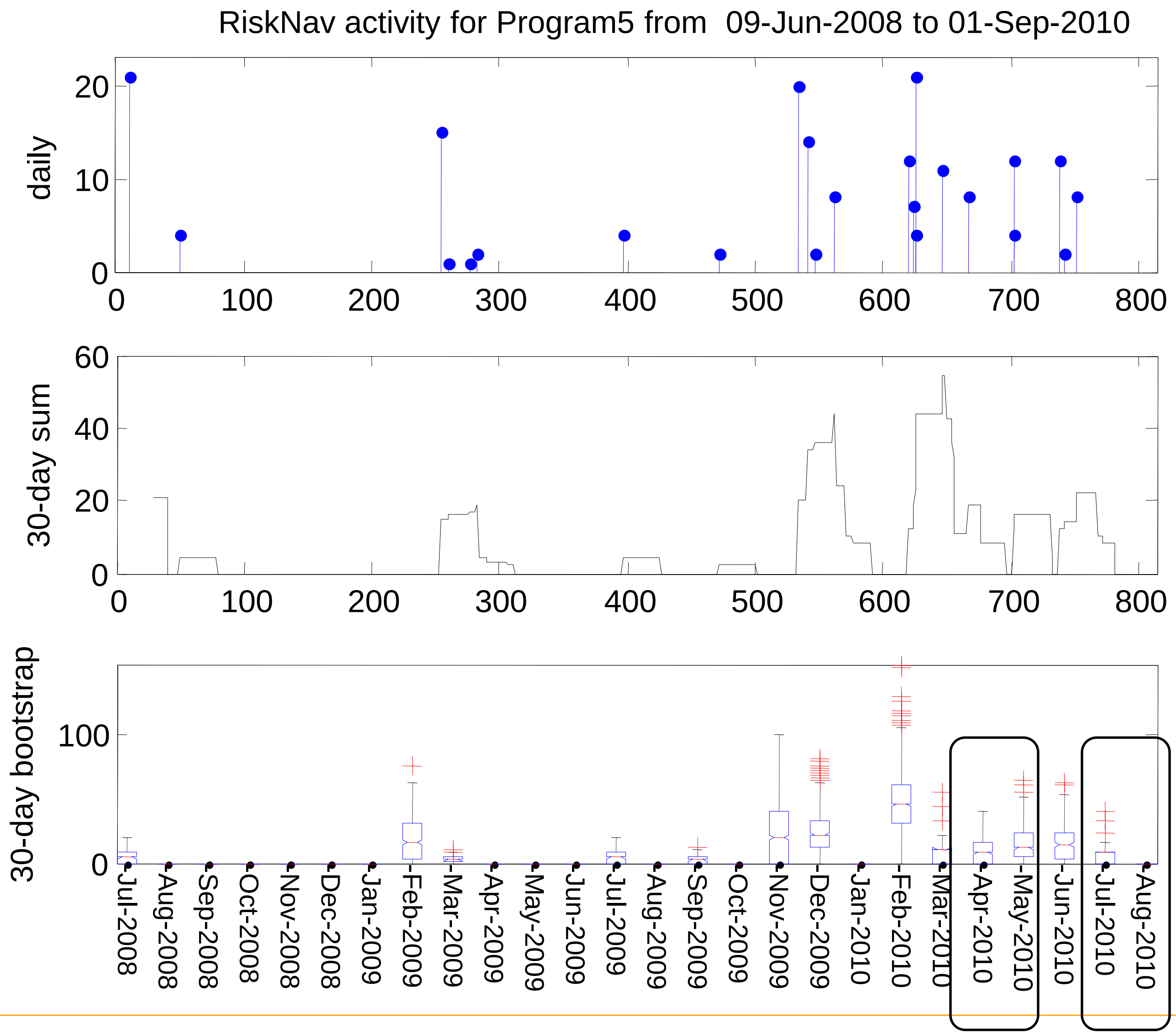
1 June 2010

- Special Note:
  - This doesn’t analyze the prediction market itself; we just assume that one took place beginning June 2010
  - This analyzes project risk-management activity, comparing 2 months before with 2 months after the prediction market began

Database activity over time

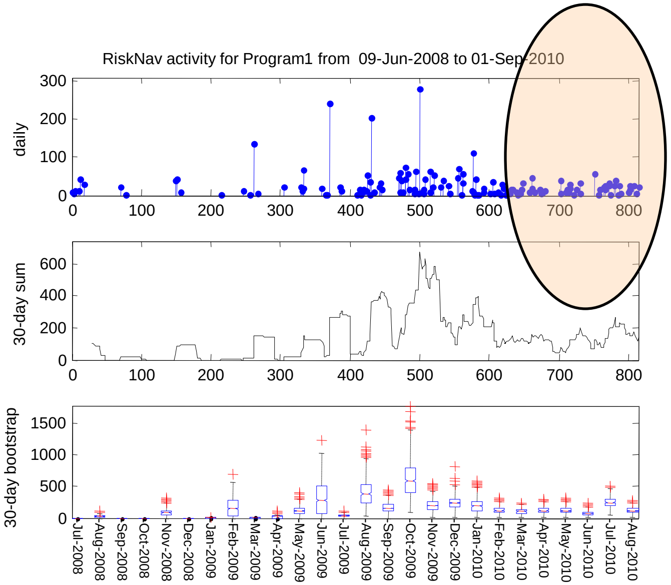


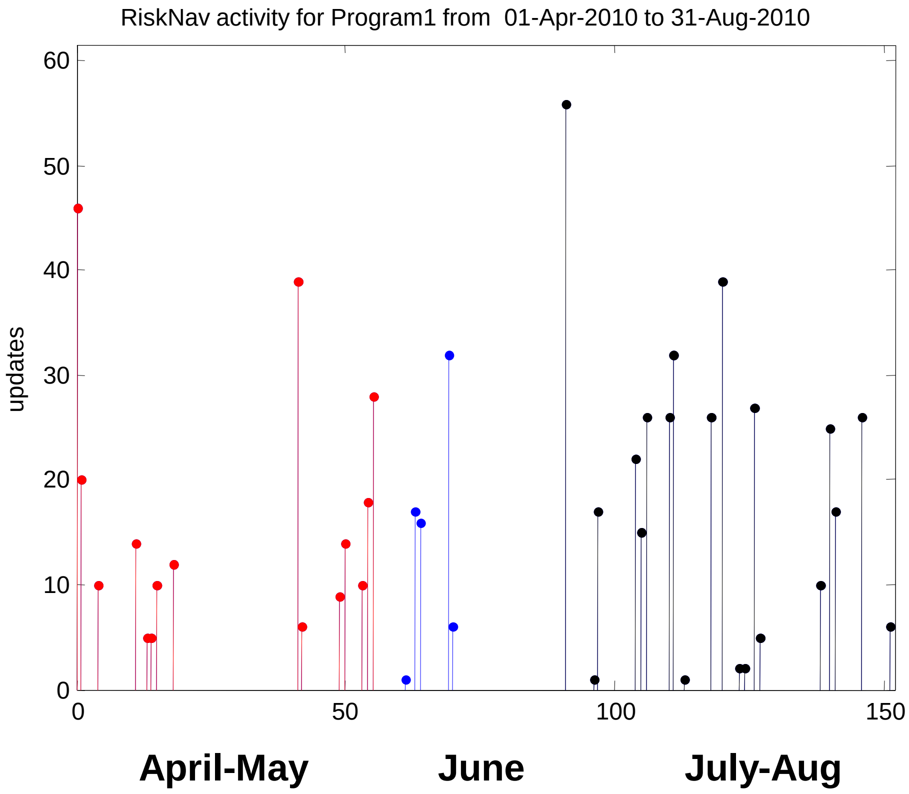
Database activity over time



Program1 update activity

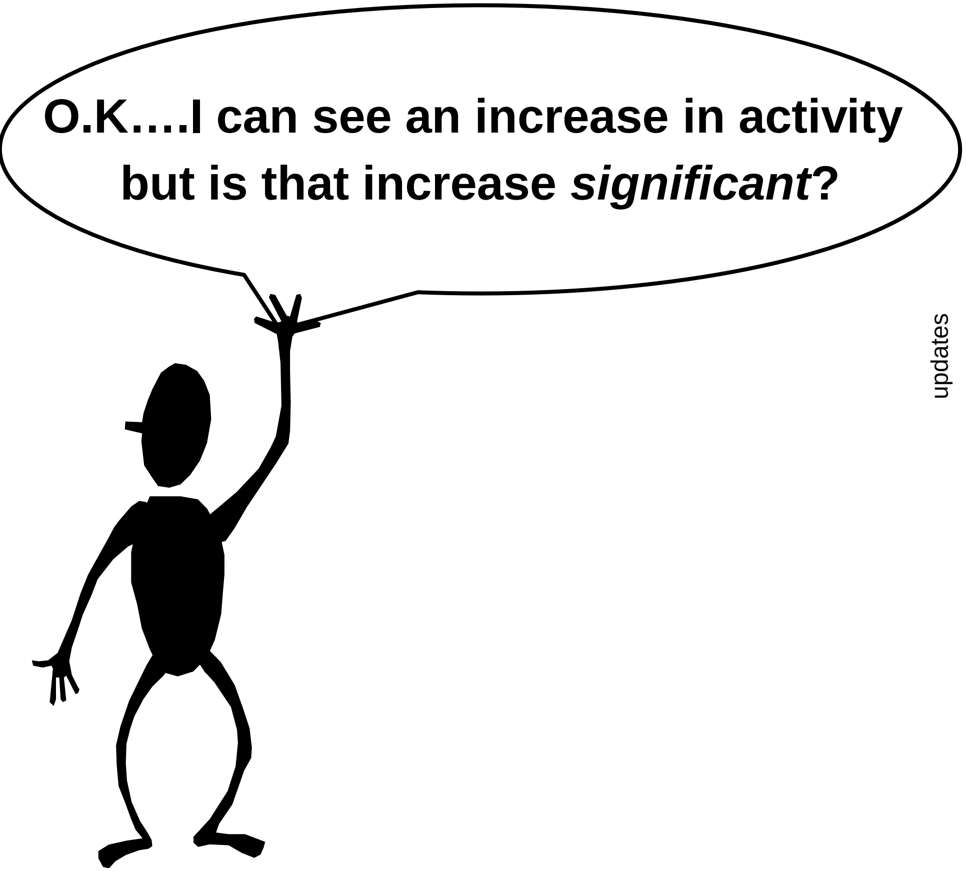
- **April-May updates: 246 (Baseline)**
- **July-Aug updates: 381 (Test period)**

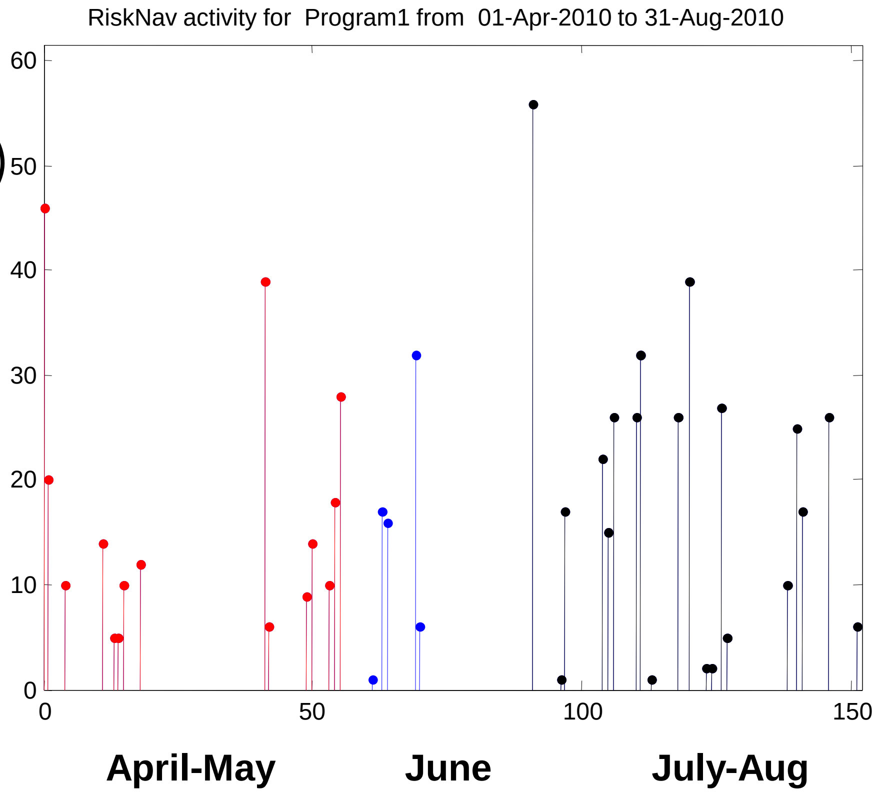




Program1 update activity

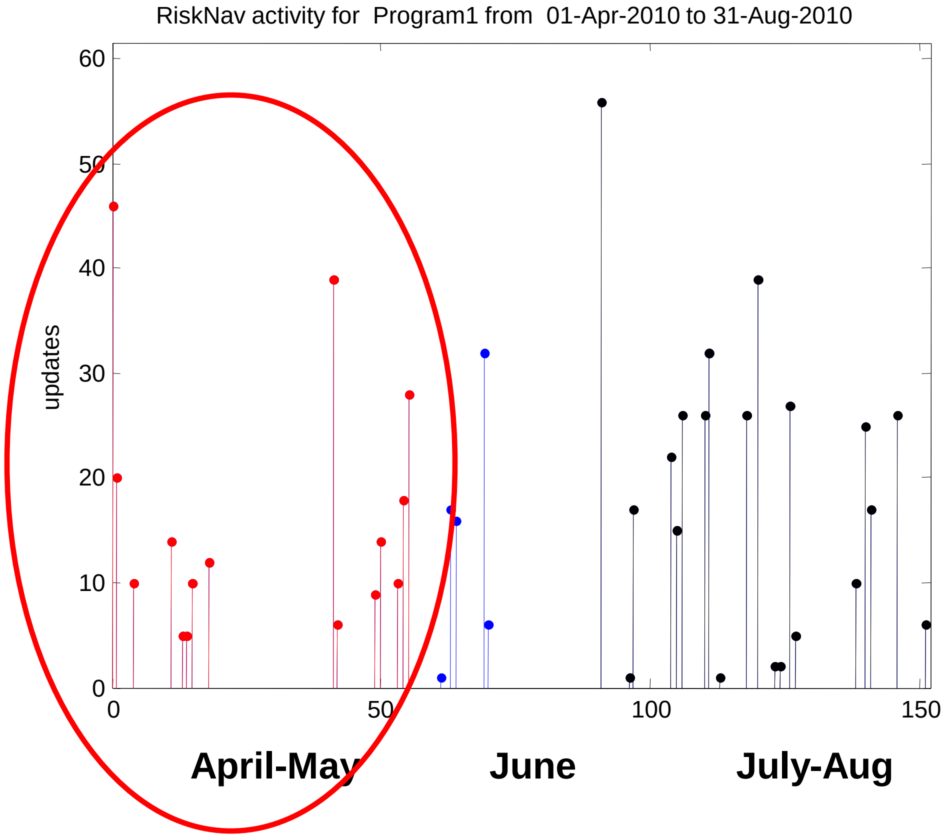
- April-May updates: 246   (Baseline)
- July-Aug updates: 381   (Test period)


O.K….I can see an increase in activity
but is that increase *significant*?



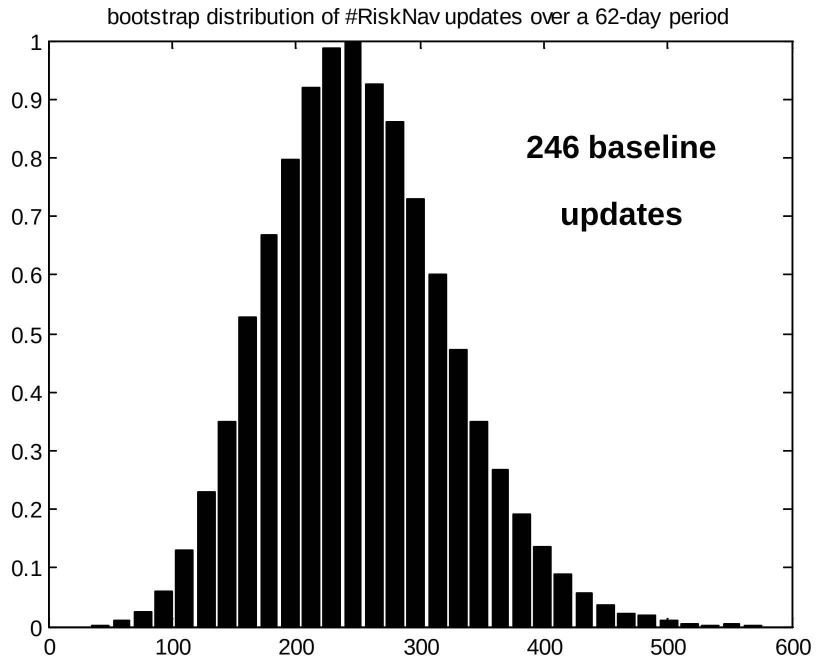
Bootstrap re-sampling of baseline (Apr-May) data

- Pick 62 days at random (with replacement) from baseline period
  - Tally up the total number of database updates observed in re-sample



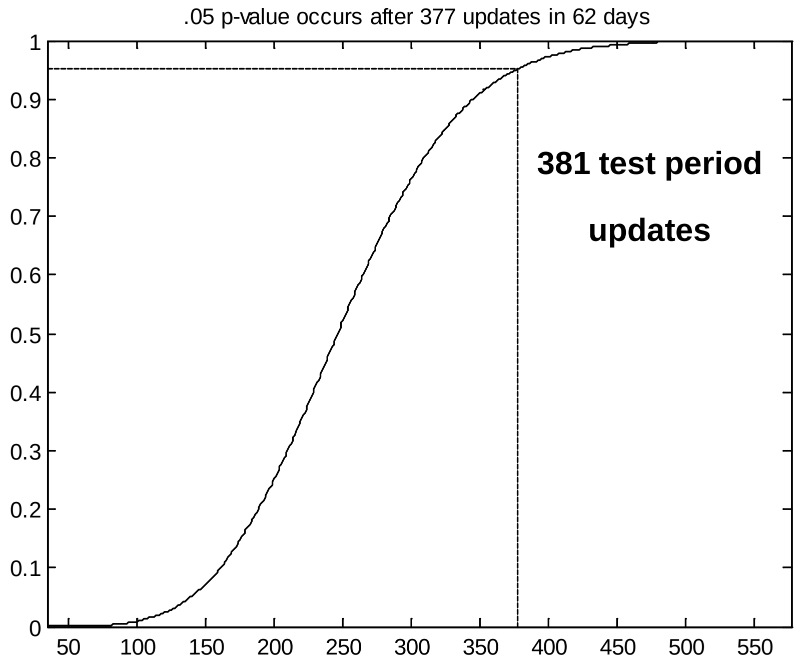
Bootstrap re-sampling of baseline (Apr-May) data

- Pick 62 days at random (with replacement) from baseline period
  - Tally up the total number of database updates observed in re-sample
- Repeat a few thousand times; build empirical distribution



bootstrap distribution of #RiskNav updates over a 62-day period

246 baseline updates



.05 p-value occurs after 377 updates in 62 days

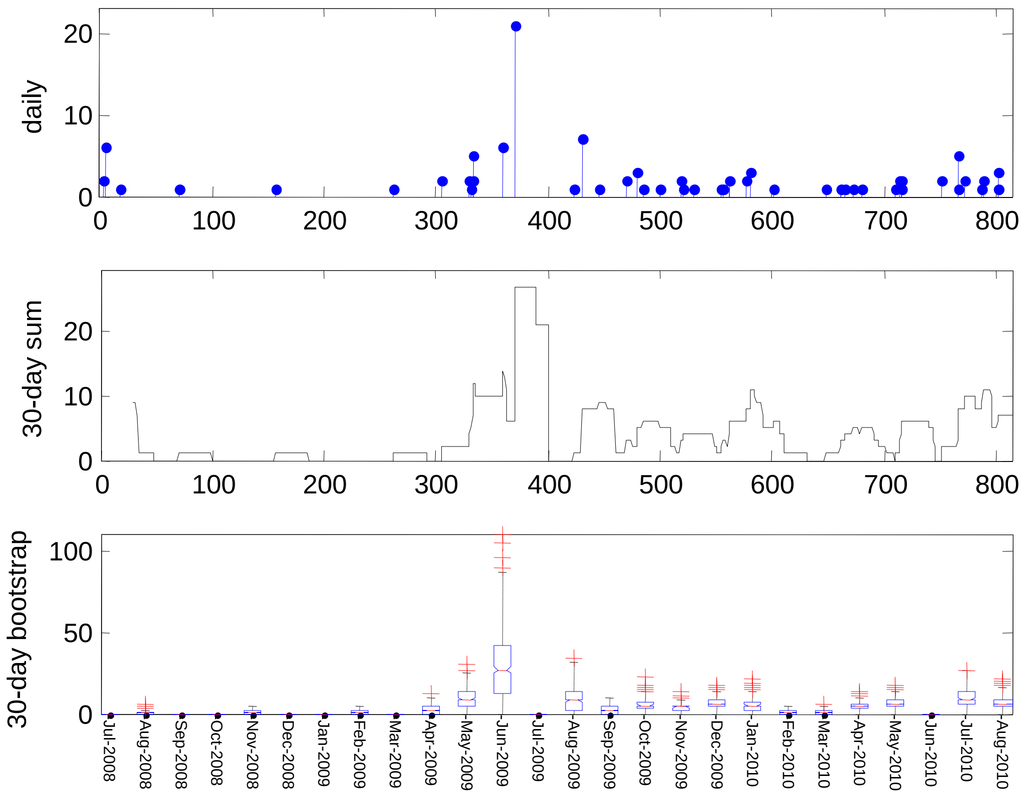
381 test period updates

- 1-sided hypothesis test
  - If we see 377 or more RiskNav updates in 62 days, we can ascribe this increase to the Prediction Market…at the 5% chance this could have otherwise occurred ‘naturally’

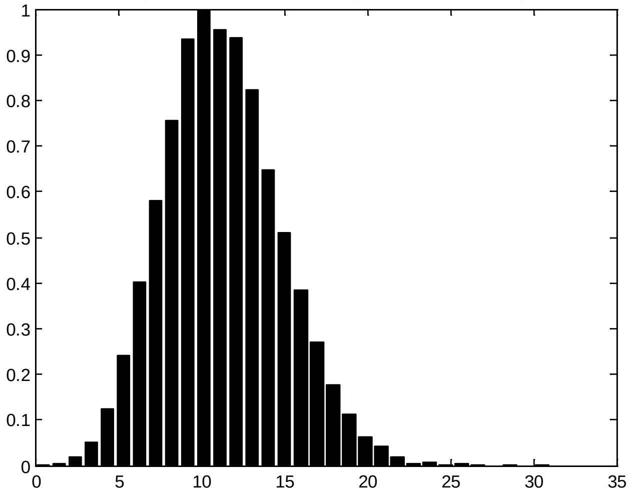
Examine the rate of newly identified risks

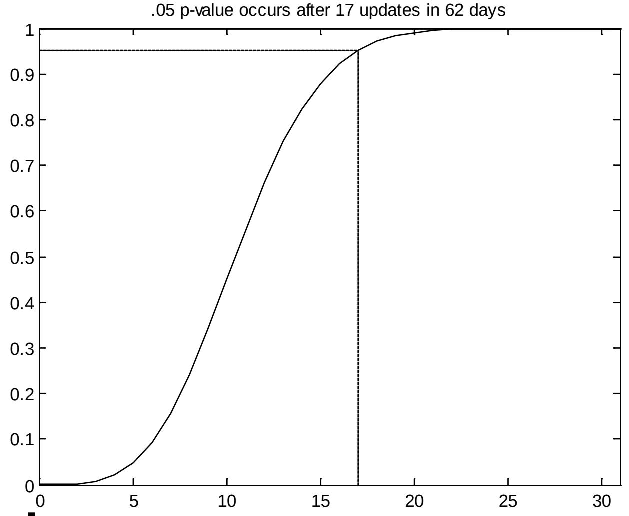
- **Corresponding bootstrap analysis: how many new risks must we see in a 62-day period?**

New Risks for Program1 from 09-Jun-2008 to 01-Sep-2010



bootstrap distribution of #RiskNav updates over a 62-day period

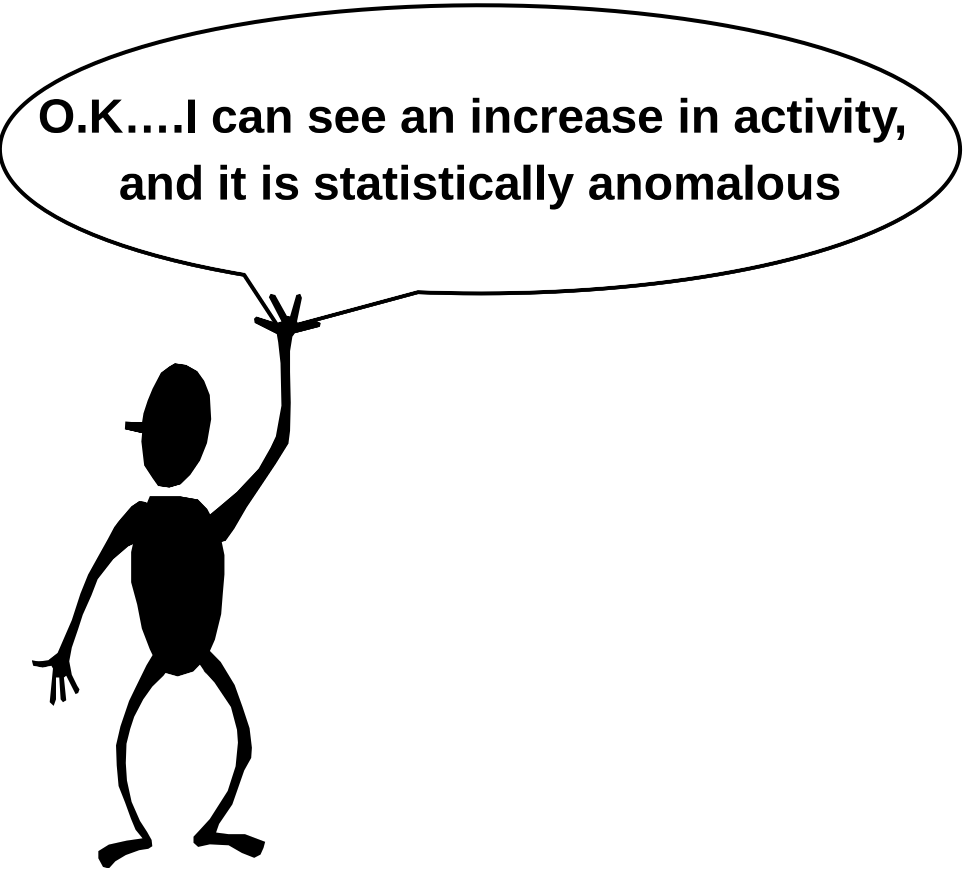


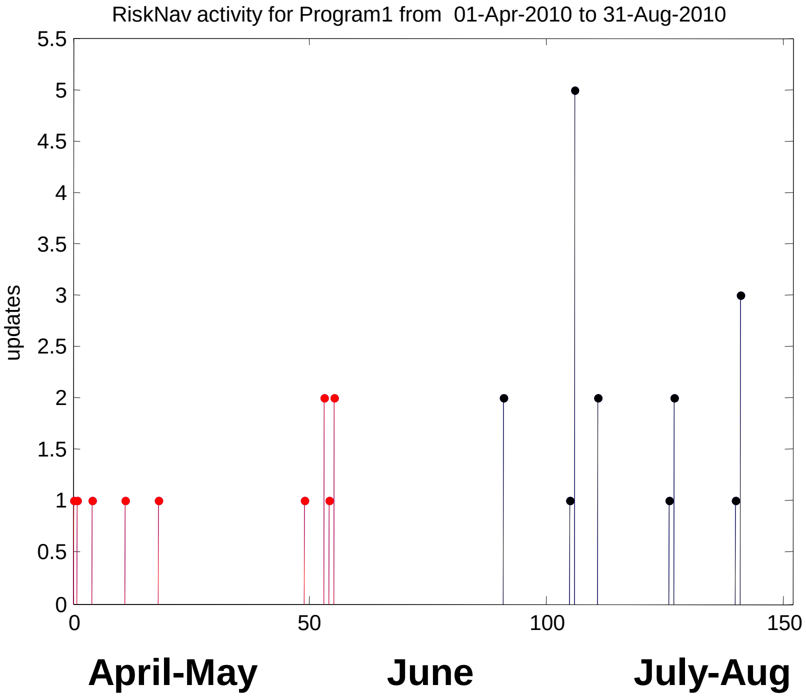


- **Concluding answer: need to see 17+ new risks**

Program1 new risks

- April-May new risks: 11 (Baseline)
- July-Aug new risks: 17 (Test period)


O.K….I can see an increase in activity, and it is statistically anomalous

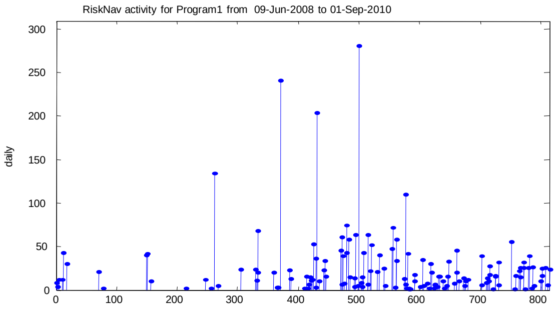


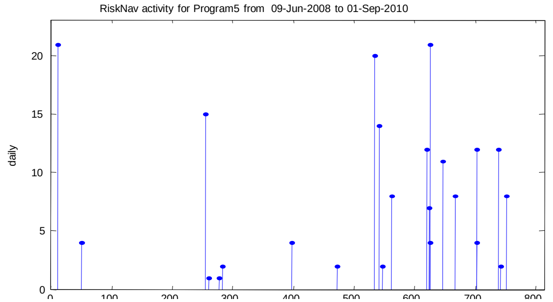
Update & New Risk activity for all programs

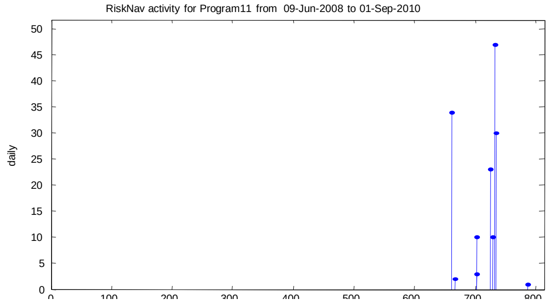
Risk Management		Update	New	# Observed Updates				# Observed New Risks			
Process Maturity	Program	Activity	Risks	Baseline	Test	.05 Level	p-value	Baseline	Test	.05 Level	p-value
Self-Generating	Program1	Y	Y	246	381	376	0.044	11	17	17	0.047
	Program2	*	Y	258	117	377	0.991	6	8	6	0.003
	Program3	**		162	145	246	0.635	3	0	3	0.640
Stimulus Driven	Program4	Y	Y	100	222	173	0.003	0	5	0	0.000
	Program5			24	8	52	0.839	1	0	3	0.646
	Program6	Y		8	21	20	0.027	3	0	0	0.000
	Program7			63	56	132	0.541	2	1	7	0.738
	Program8	Y	Y	10	30	24	0.012	0	2	0	0.000
Initiating	Program9			0	0	0	0.000	0	0	0	
	Program10		Y	0	20	0	0.000	0	3	0	0.000
	Program11			46	1	117	0.967	3	0	7	0.870
Inactive	Program12			0	0	0		0	0	0	
	Program13			0	0	0		0	0	0	
	Program14			0	0	0		0	0	0	

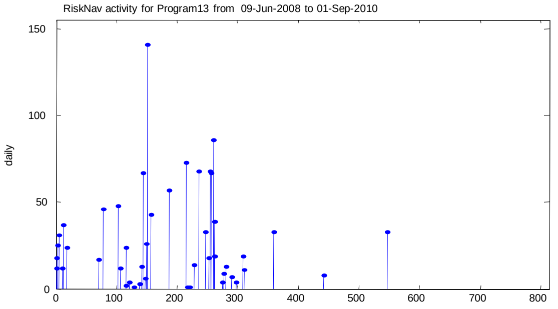
Let’s break this down into something simpler

- **4 types of risk-management cultures observed in the 14 programs**
  - **Self-generating -** routine updates on a consistent basis with no evident stimulus
  - **Stimulus Driven –** periodic updates as the result of external events (e.g. PMRs, risk meetings, risk team interactions)
  - **Initiating –** programs initializing a risk management process or use of the tool
  - **Inactive –** programs no longer actively managing risks



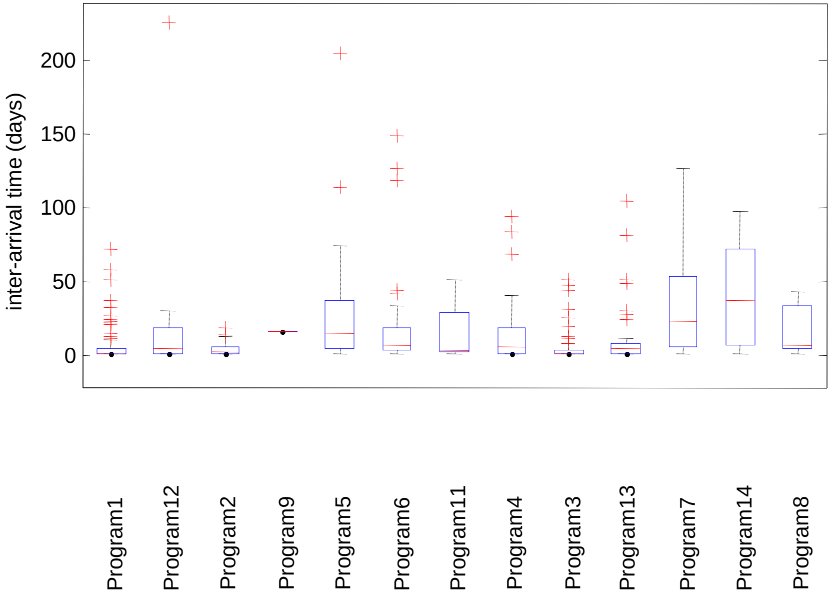


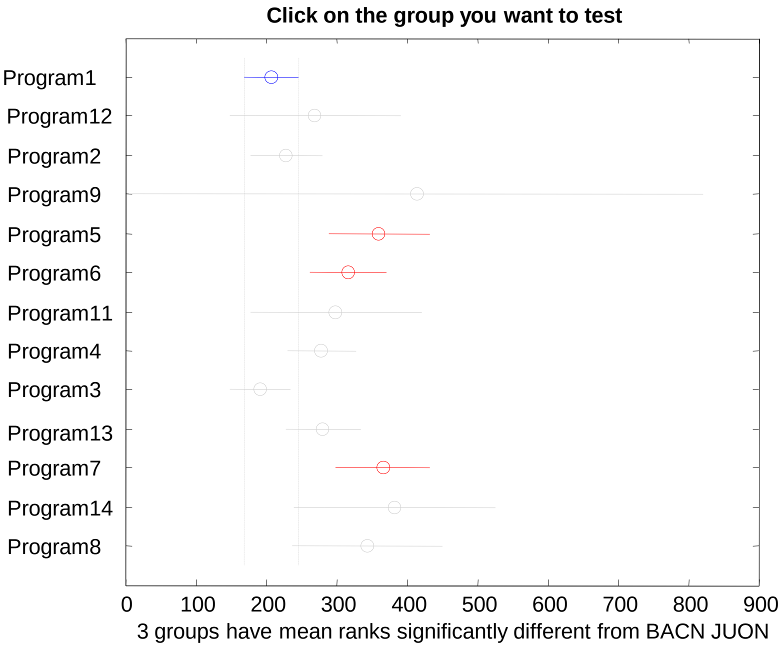




You can distinguish programs rigorously…

- Look at the #days between database entries
- Kruskal-Wallace rank-based 1-way ANOVA tells demarks programs by significantly different median inter-arrival times





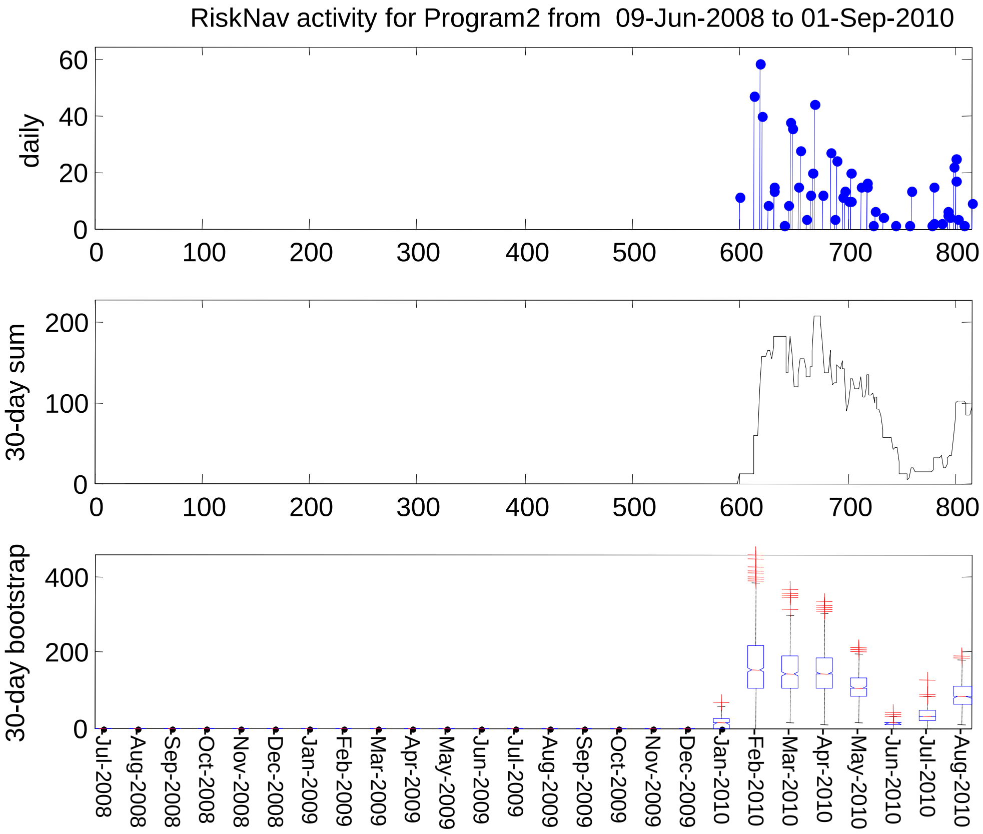
Actionable summary

- Treatment effect concentrated in programs with self-generating risk dialogues
  - *Program2: just coming on-line to RiskNav
  - **Program3: Experienced program cuts in Aug.
- Treatment effect in stimulus driven is observed but discounted
  - Sufficiently large variability in sparse updates

Risk Management		Update	New
Process Maturity	**Program**	**Activity**	**Risks**
Self-Generating	Program1	Y	Y
	Program2	*	Y
	Program3	**	
Stimulus-Driven	Program4	Y	Y
	Program5		
	Program6	Y	
	Program7		
	Program8	Y	Y
Initiating	Program9		
	Program10		Y
	Program11		
Inactive	Program12		
	Program13		
	Program14		

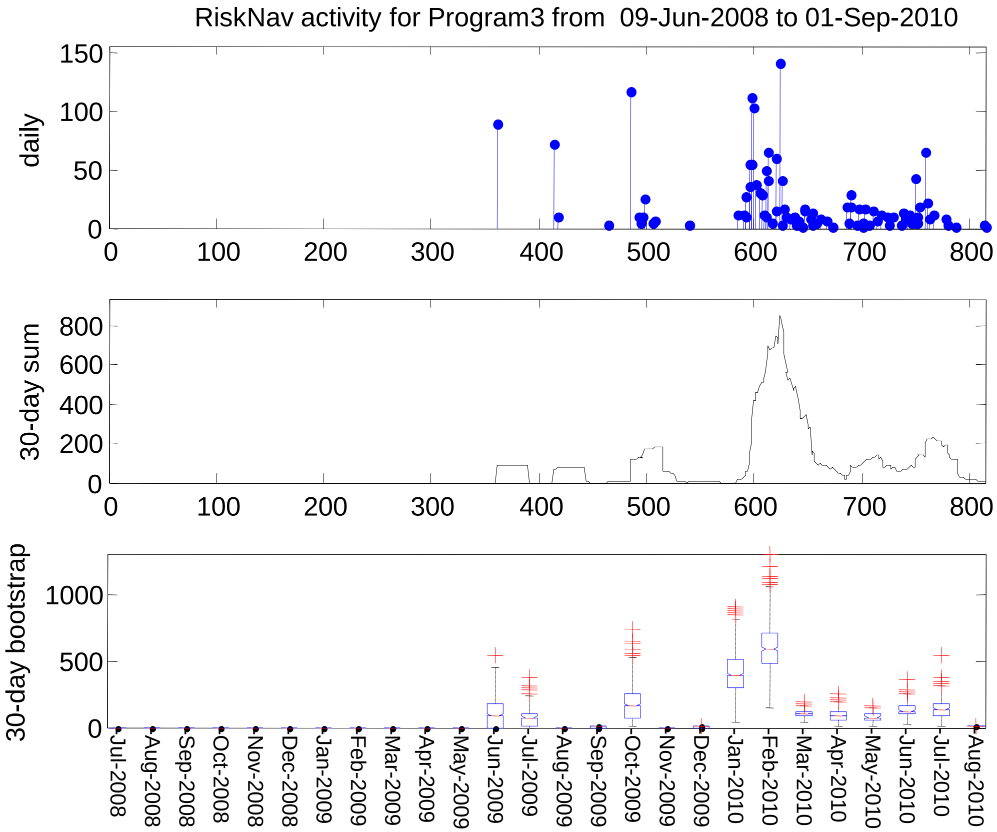
Program2 in more detail

- Program transitioning into RiskNav tool
- Abnormally high activity during baseline period.
- Don’t see a treatment effect during test period in update activity
- Do see a treatment effect for new risks



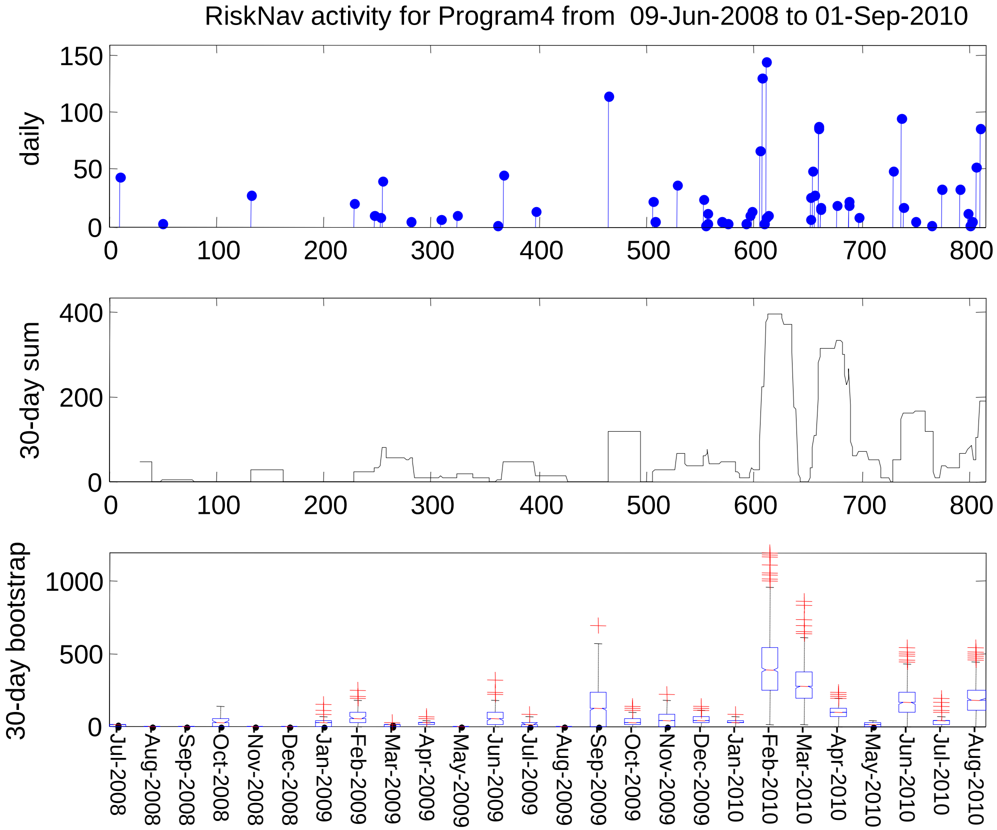
Program3 in more detail

- Mission fundamentally re-evaluated during test period
  - Program cuts August 2010
- Treatment effect is observed comparing {July} vs. {April-May}

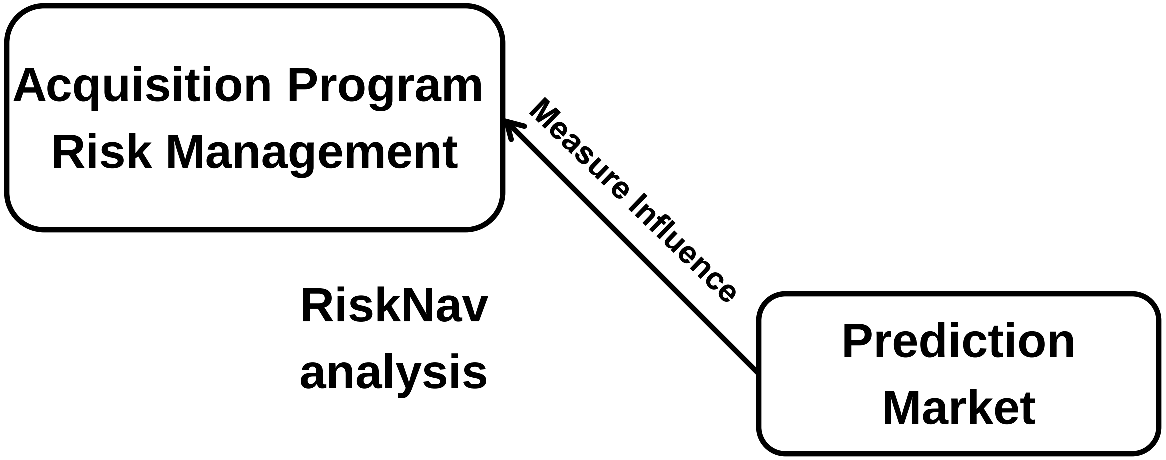


Program4 in more detail

- 100+ update day occurred in June
- Had that stimulus-day occurred in May or July, would have come to different conclusions
- Need longer observation time to declare/deny treatment effect


RiskNav activity for Program4 from 09-Jun-2008 to 01-Sep-2010

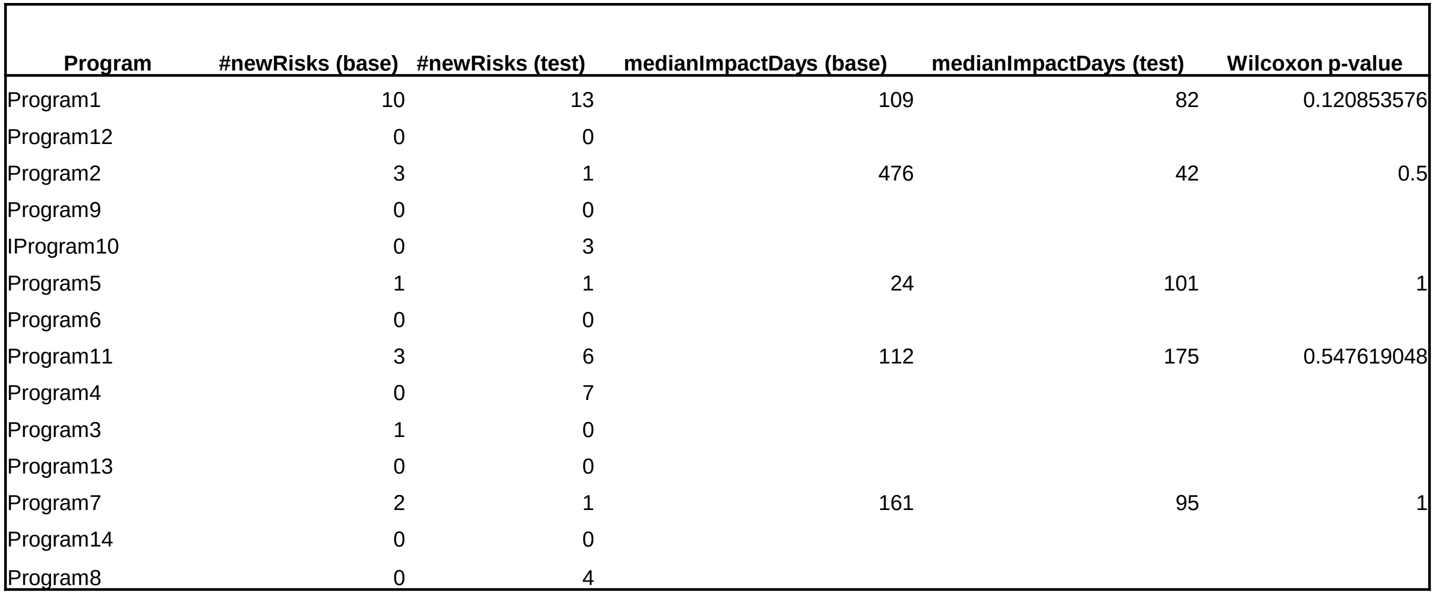
‘Treatment Effects’ we would expect to see



- Is there an increase in the overall database activity?
- Is there an increase the rate of newly identified risks?
- Are new risks identified earlier from their event-horizon?
- Do risks get mitigated or closed more quickly?

Risk lead time

- **Subset of new risks specified a lead time:**
- **Wilcoxon rank-sum test: Do two populations have equally large values?**
  - **Insufficient data to draw any conclusion**

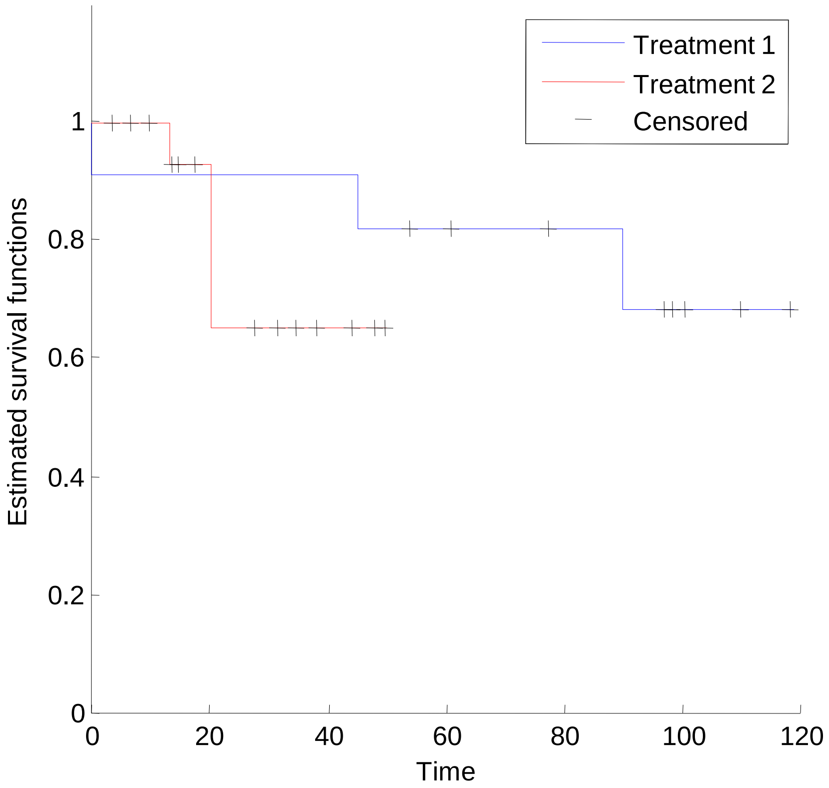


Program	#newRisks (base)	#newRisks (test)	medianImpactDays (base)	medianImpactDays (test)	Wilcoxon p-value
Program1	10	13	109	82	0.120853576
Program12	0	0			
Program2	3	1	476	42	0.5
Program9	0	0			
IProgram10	0	3			
Program5	1	1	24	101	1
Program6	0	0			
Program11	3	6	112	175	0.547619048
Program4	0	7			
Program3	1	0			
Program13	0	0			
Program7	2	1	161	95	1
Program14	0	0			
Program8	0	4			

Risk closures

- Constructed Kaplan-Mier survival curves for new risks
- Log-Rank test determines if one population survives longer than another
  - Insufficient data to declare closure times are shorter

LogRank test for difference in Kaplan-Meier survival function for Program1 between base and test period



Possible concerns with this analysis

- Seasonal adjustments
  - Insufficient RiskNav histories to make meaningful adjustments
  - Traditionally, July-August is a period of reduced work activity
    - We observe an increase in risk-management activity despite this
- Connection with the Hawthorne effect
  - This analysis doesn’t focus on specific questions or derived information
  - Improvements may be result of general employee feedback in a prediction market
- This work was exploratory research, not a randomized clinical trial
  - We cannot statistically rule out some other (non-prediction market) effect going on here
- However
  - This provides strong direction for future research inquiries
  - Analogous to passing ‘Phase-0’ trials in FDA approval process

Conclusions

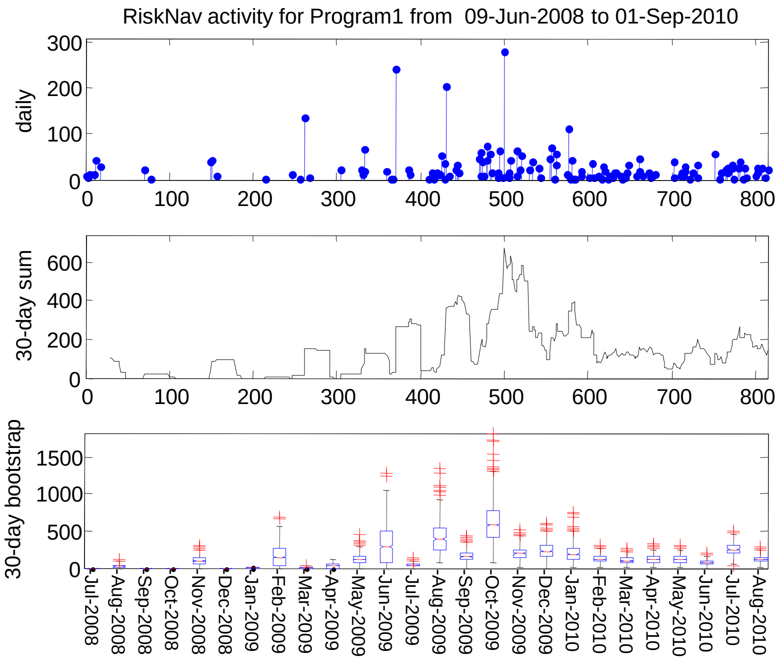
- After introducing a prediction market to the USAF…
  - We observe enhanced risk management practices on programs with self-generating risk identification and mitigation activity
  - We do not conclusively observe this effect on programs with stimulus-driven risk management processes
- There was not enough data to evaluate whether prediction markets would foster earlier risk identification or faster closure of risks

Update Activity and New Risks

Program Analysis Data

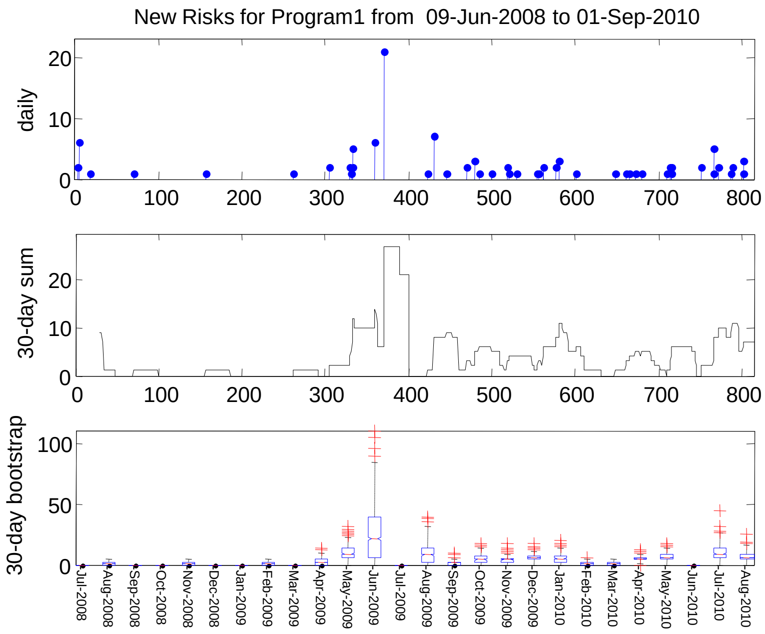
Program1

Update Activity



# Observed Base	246
# Observed Test	381
# at .05 Level	376
p-value	.044

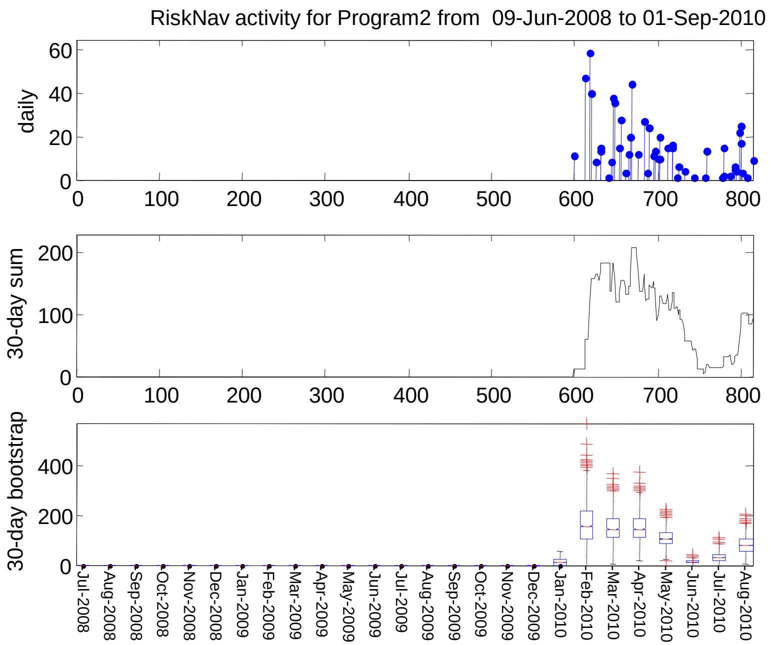
New Risks Identified



# Observed Base	11
# Observed Test	17
# at .05 Level	17
p-value	.047

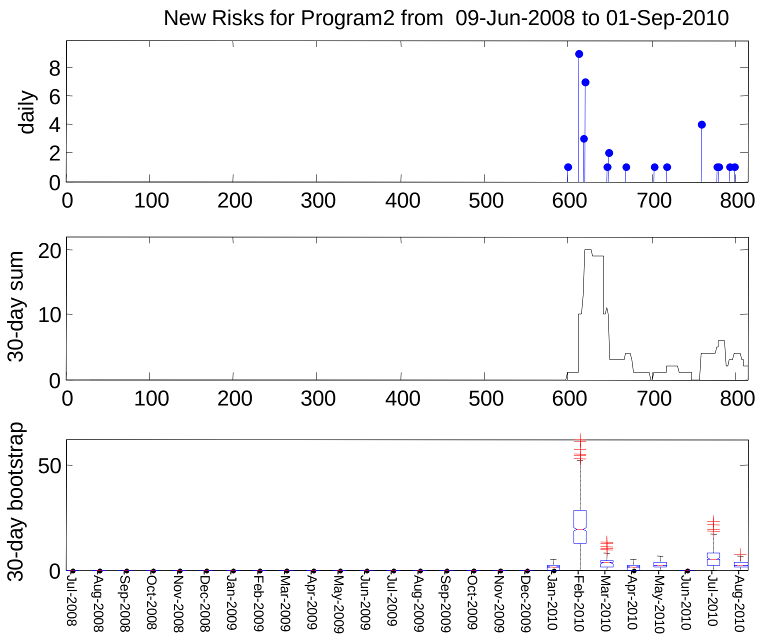
Program2

Update Activity



# Observed Base	258
# Observed Test	177
# at .05 Level	377
p-value	.991

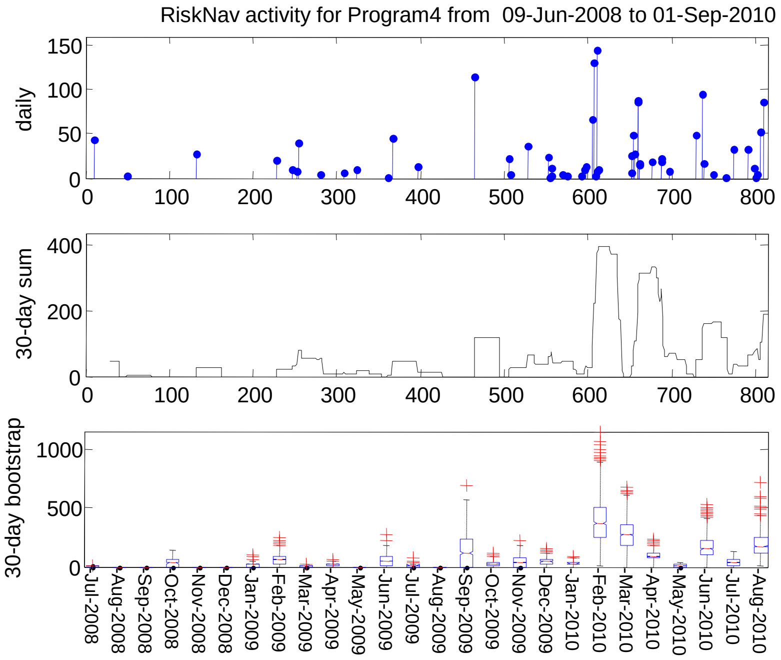
New Risks Identified



# Observed Base	3
# Observed Test	8
# at .05 Level	6
p-value	.003

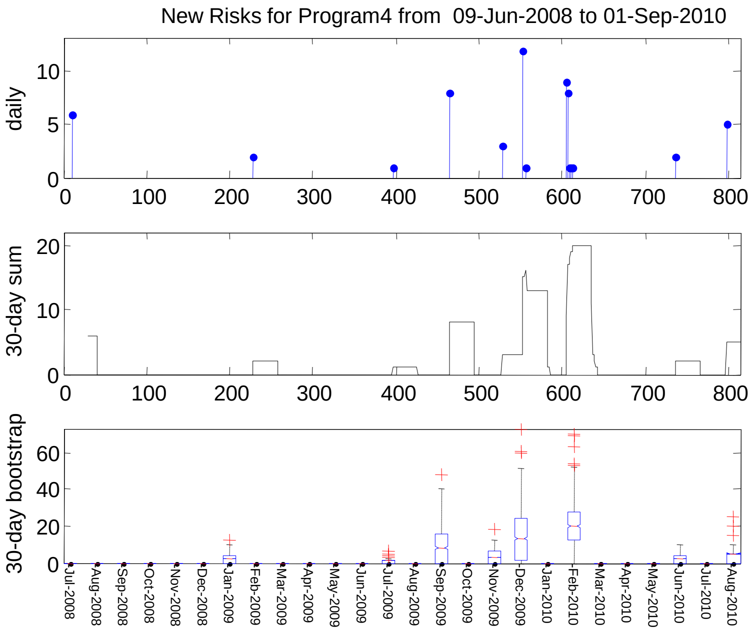
Program4

Update Activity



# Observed Base	100
# Observed Test	222
# at .05 Level	173
p-value	.003

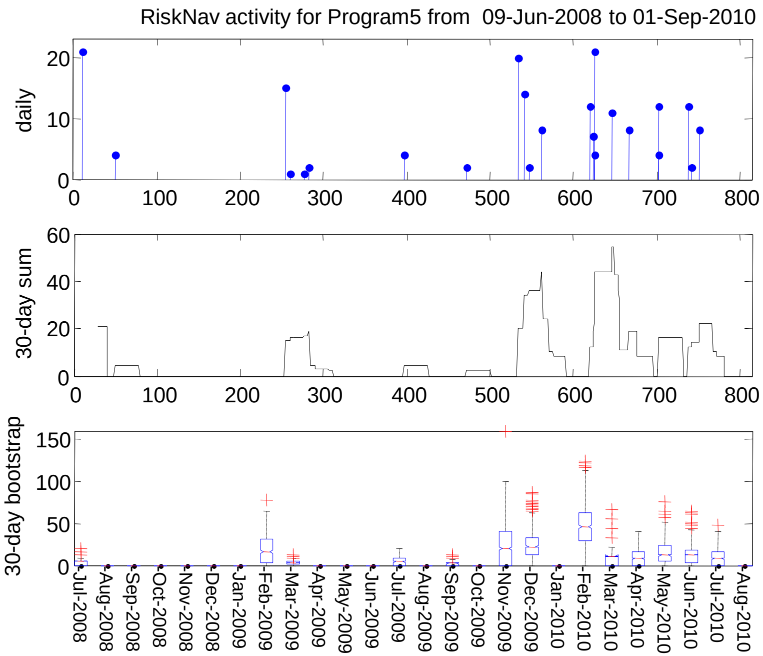
New Risks Identified



# Observed Base	0
# Observed Test	5
# at .05 Level	0
p-value	.000

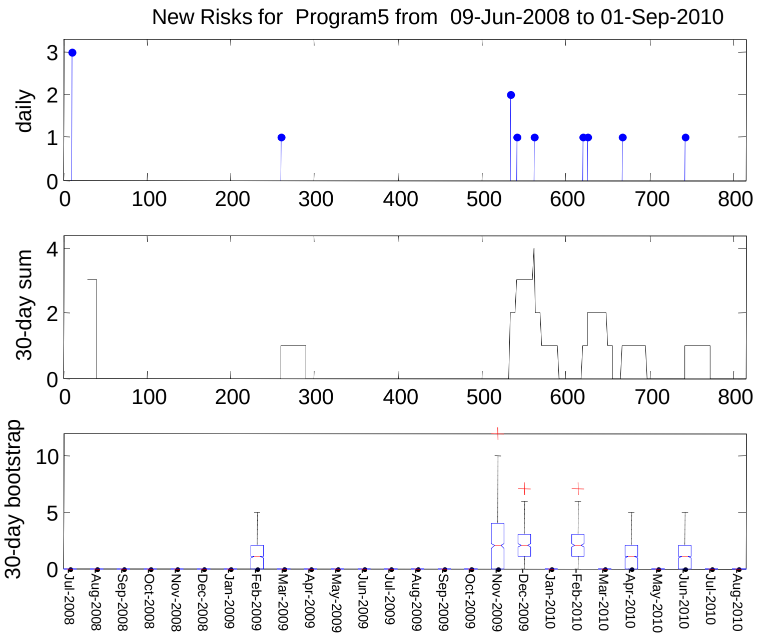
Program5

Update Activity



# Observed Base	24
# Observed Test	8
# at .05 Level	52
p-value	.839

New Risks Identified

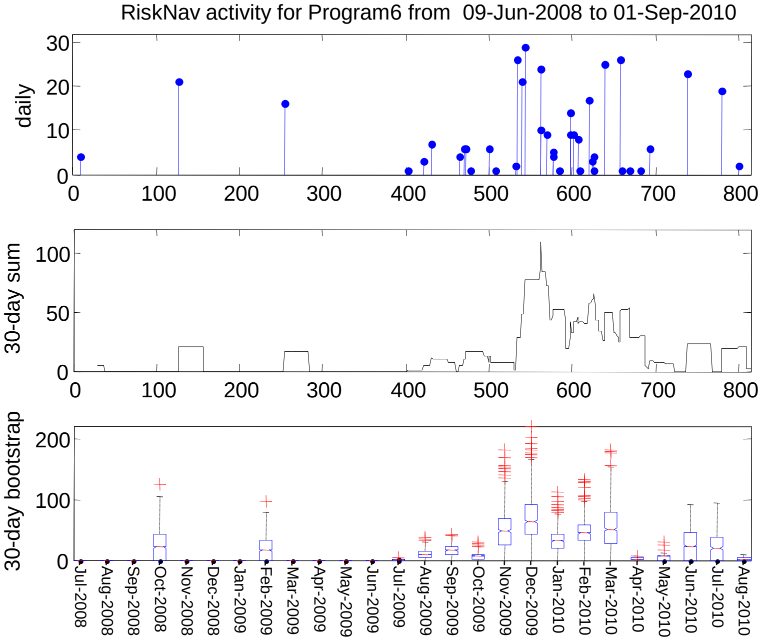


# Observed Base	1
# Observed Test	0
# at .05 Level	3
p-value	..636

MITRE

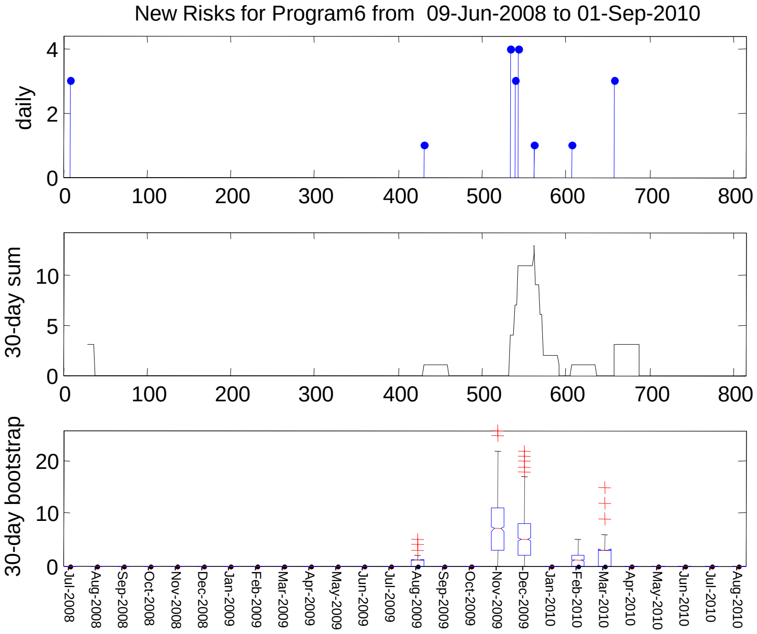
Program6

Update Activity



# Observed Base	8
# Observed Test	21
# at .05 Level	20
p-value	.027

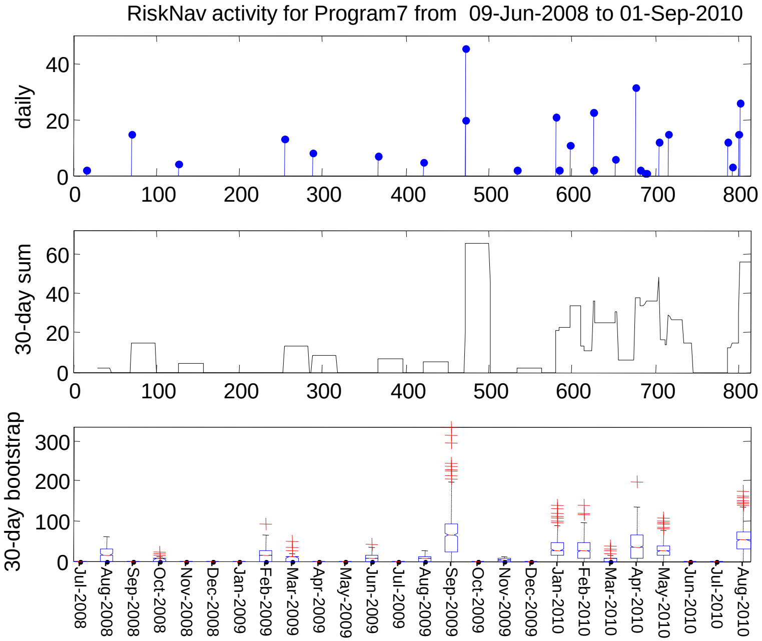
New Risks Identified



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

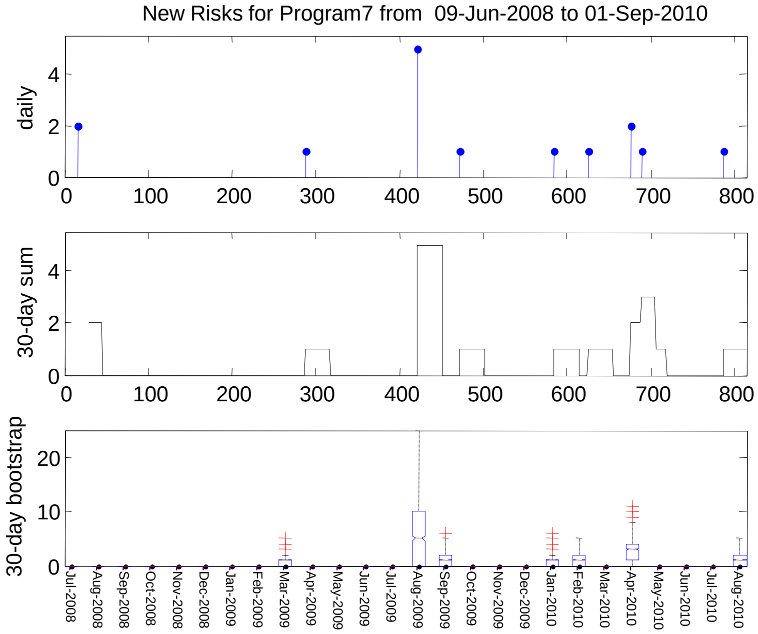
Program7

Update Activity



# Observed Base	63
# Observed Test	56
# at .05 Level	132
p-value	.541

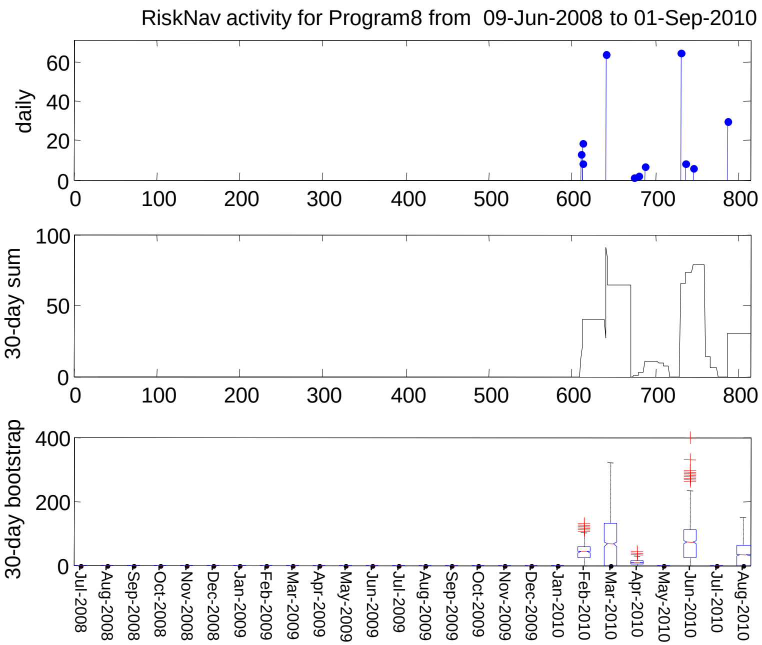
New Risks Identified



# Observed Base	3
# Observed Test	1
# at .05 Level	7
p-value	.747

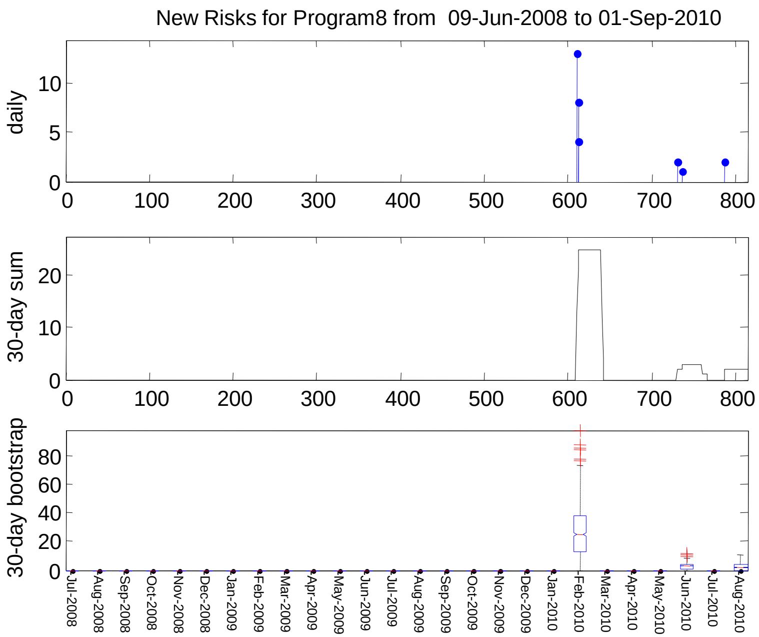
Program8

Update Activity



# Observed Base	10
# Observed Test	30
# at .05 Level	24
p-value	.012

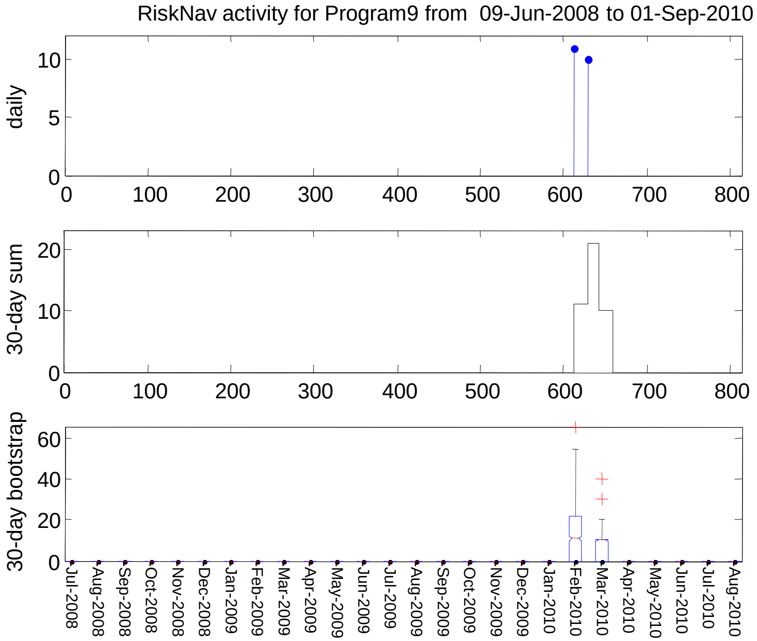
New Risks Identified



# Observed Base	0
# Observed Test	2
# at .05 Level	0
p-value	.000

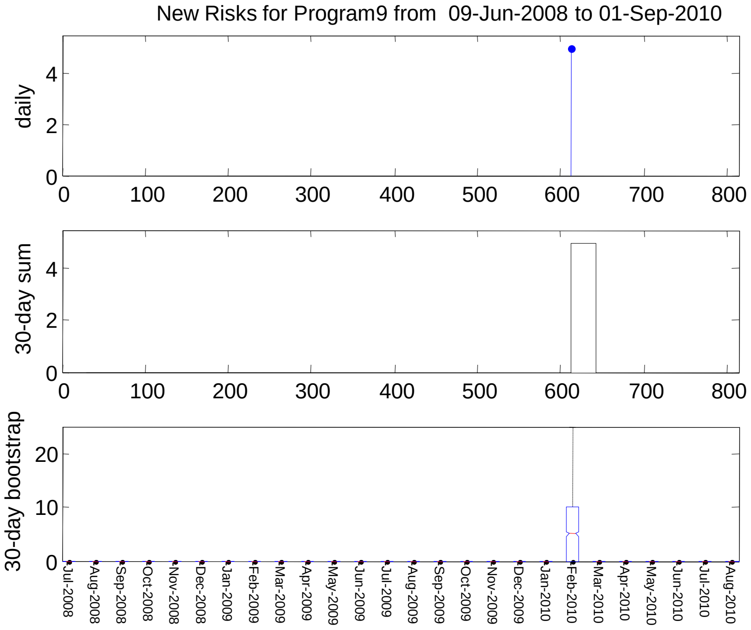
Program9

Update Activity



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

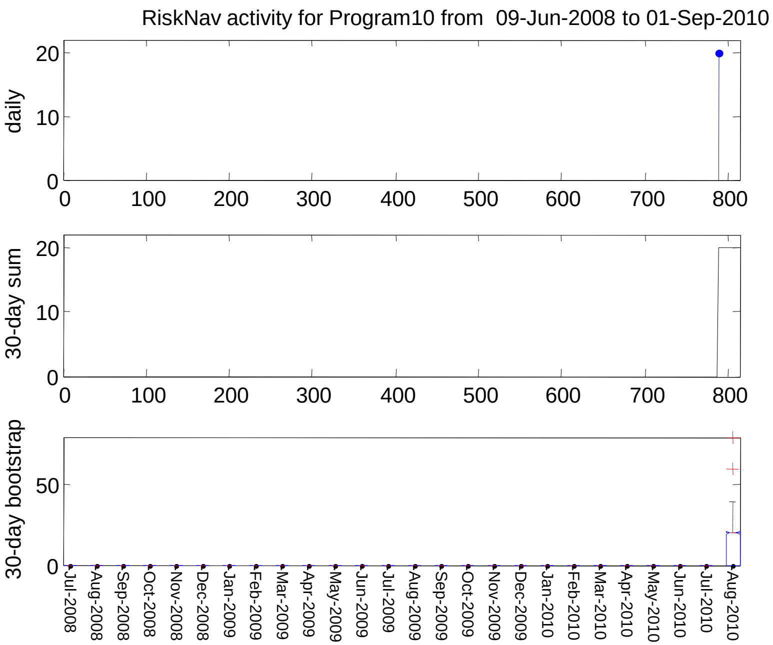
New Risks Identified



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

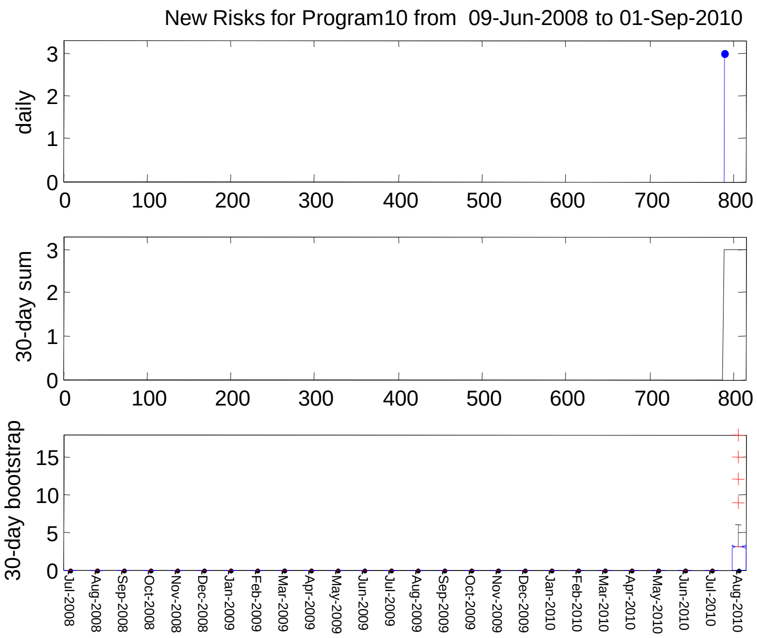
Program10

Update Activity


RiskNav activity for Program10 from 09-Jun-2008 to 01-Sep-2010

# Observed Base	0
# Observed Test	20
# at .05 Level	0
p-value	.000

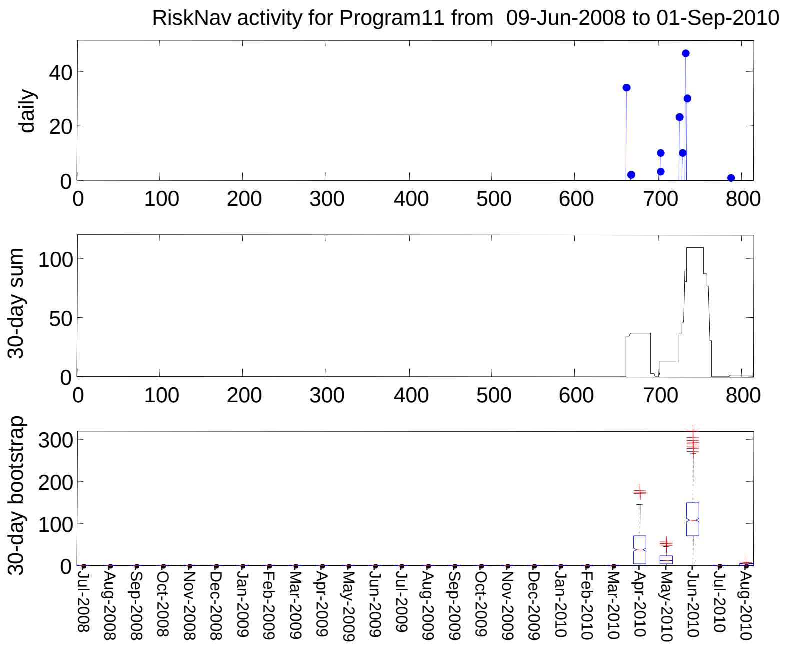
New Risks Identified


New Risks for Program10 from 09-Jun-2008 to 01-Sep-2010

# Observed Base	0
# Observed Test	3
# at .05 Level	0
p-value	.000

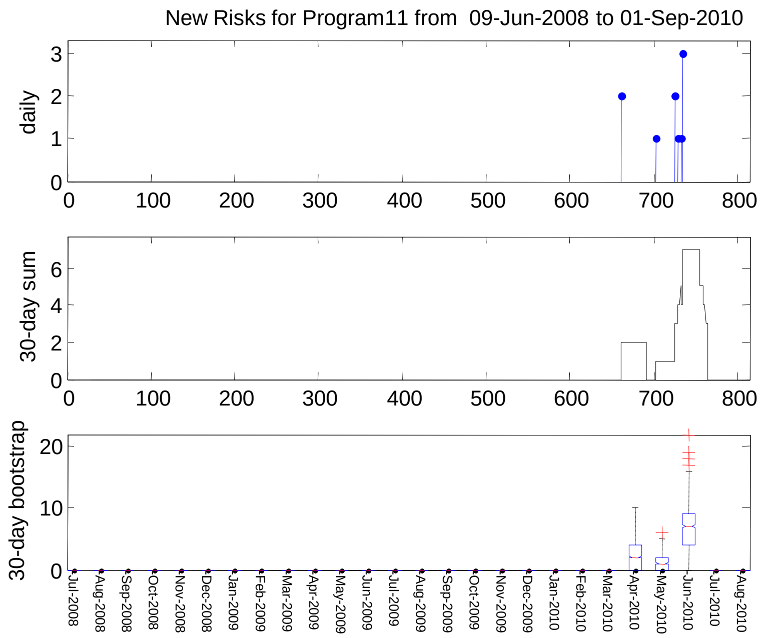
Program11

Update Activity



# Observed Base	49
# Observed Test	1
# at .05 Level	117
p-value	.967

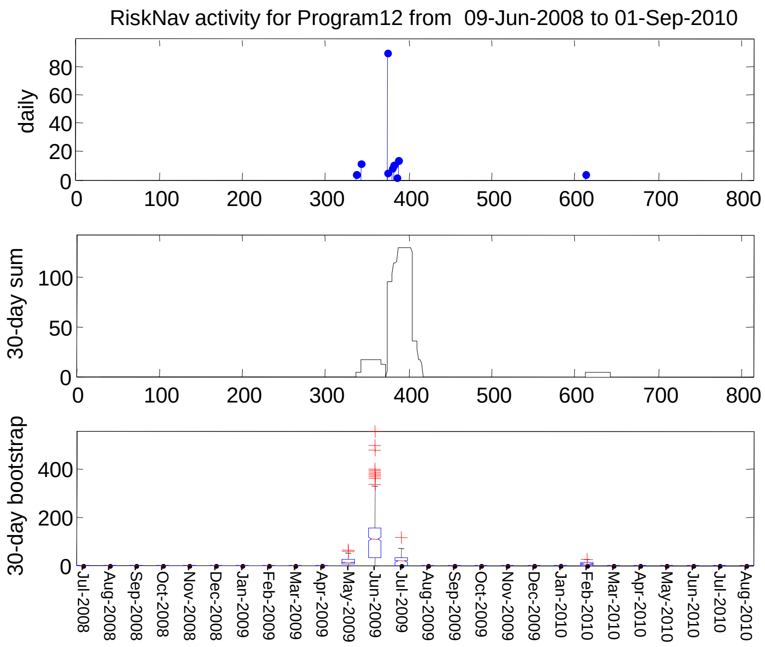
New Risks Identified



# Observed Base	3
# Observed Test	0
# at .05 Level	7
p-value	.870

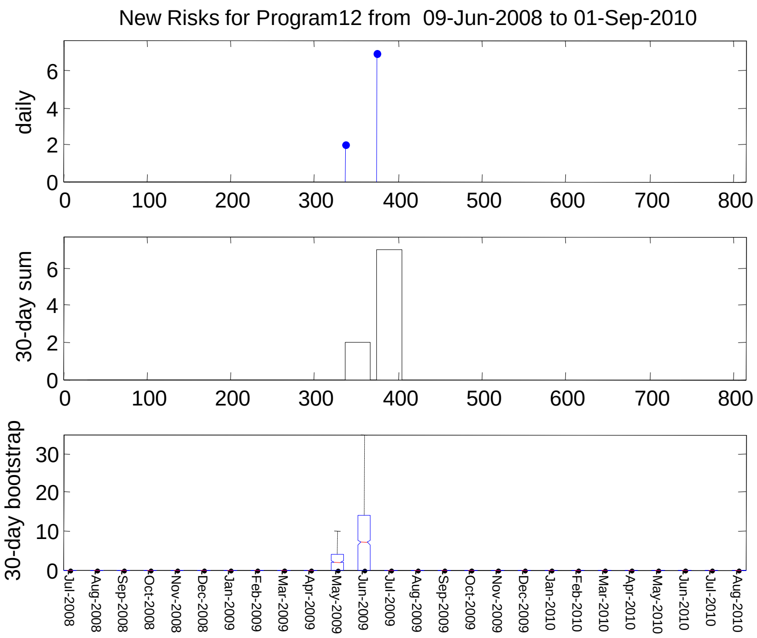
Program12

Update Activity



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

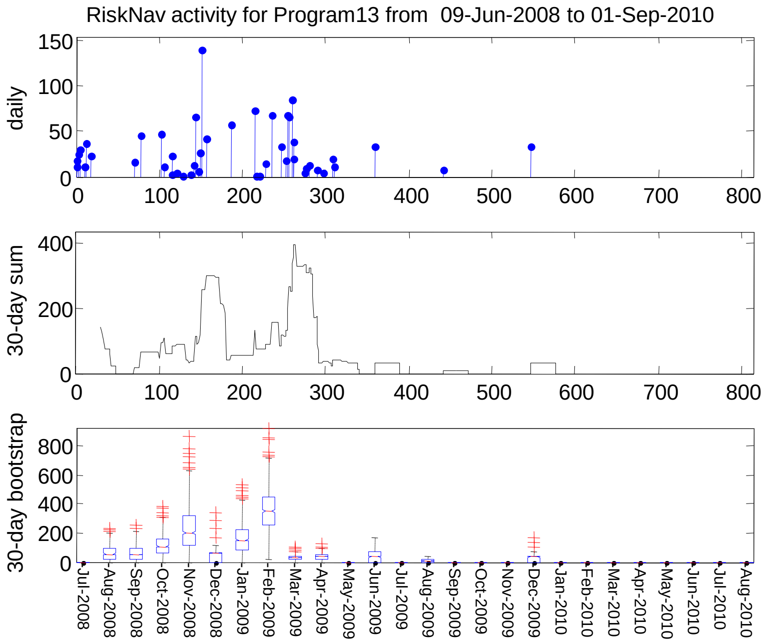
New Risks Identified



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

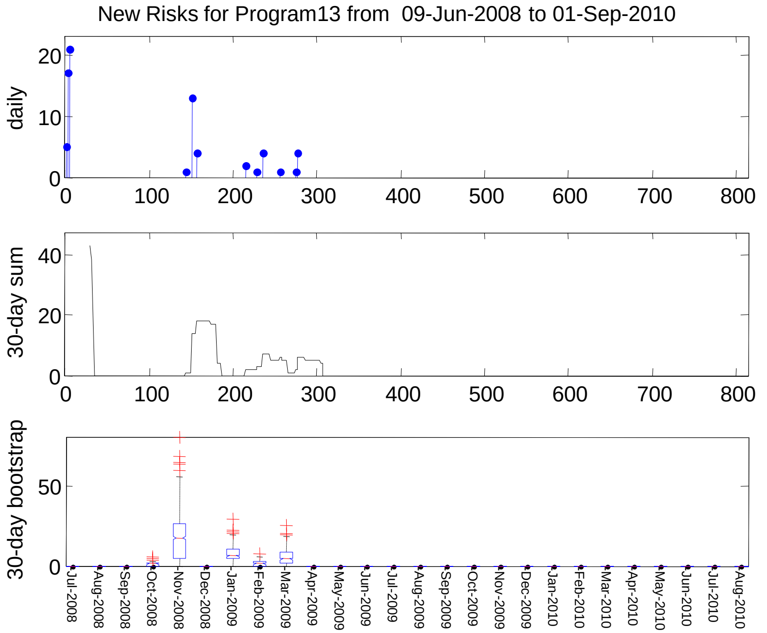
Program13

Update Activity



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

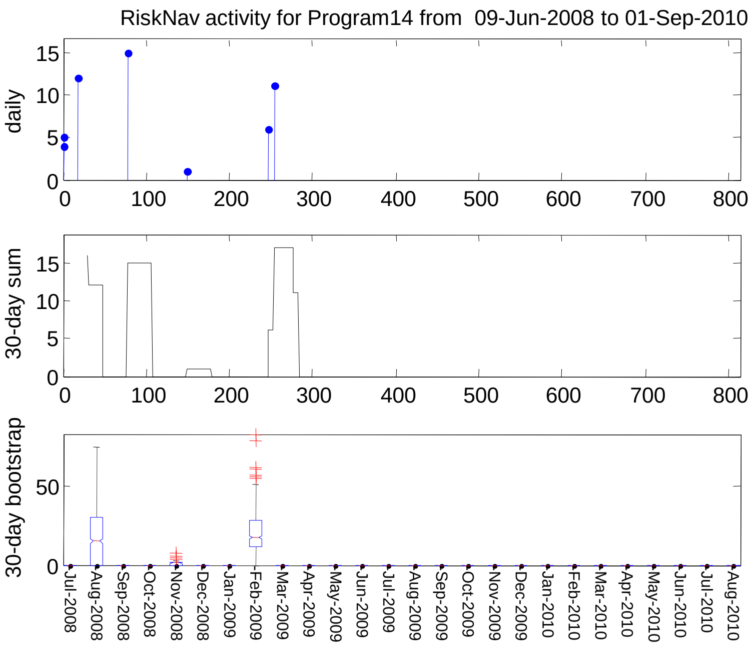
New Risks Identified



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

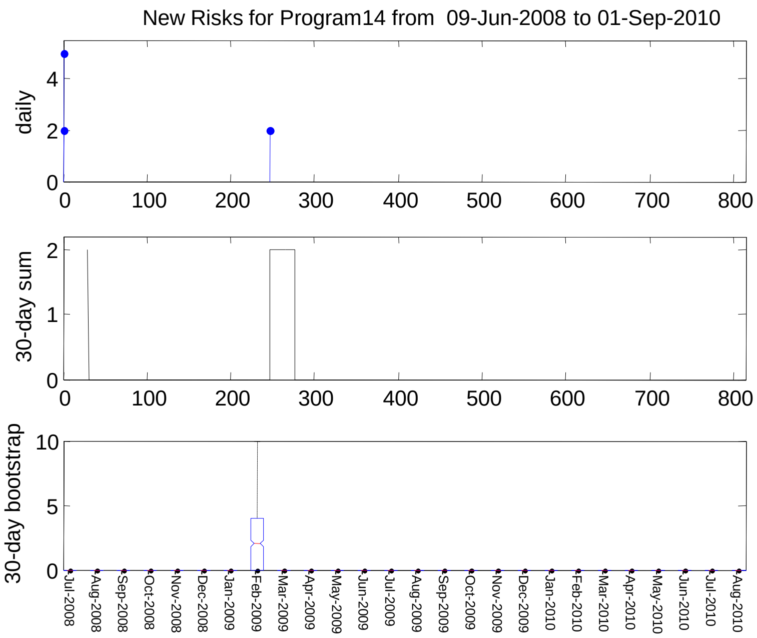
Program14

Update Activity



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	

New Risks Identified



# Observed Base	0
# Observed Test	0
# at .05 Level	0
p-value	