CCP and Prostate Cancer: Or How Not to Write Equations

Terrence P McGarty
White Paper No 98
July, 2013

We examine the elements required in modeling cancer and we propose several variants. We focus primarily on intracellular models using a differential equation approach although questioning and modifying the reaction rate models in common use. We also develop a model for total cellular dynamics which we integrate with the intracellular model thus allowing for whole body predictive capabilities.

CCP AND PROSTATE CANCER: OR HOW NOT TO WRITE EQUATIONS

Notice

This document represents the personal opinion of the author and is not meant to be in any way the offering of medical advice or otherwise. It represents solely an analysis by the author of certain data which is generally available. The author furthermore makes no representations that the data available in the referenced papers is free from error. The Author also does not represent in any manner or fashion that the documents and information contained herein can be used other than for expressing the opinions of the Author. Any use made and actions resulting directly or otherwise from any of the documents, information, analyses, or data or otherwise is the sole responsibility of the user and The Author expressly takes no liability for any direct or indirect losses, harm, damage or otherwise resulting from the use or reliance upon any of the Author's opinions as herein expressed. There is no representation by The Author, express or otherwise, that the materials contained herein are investment advice, business advice, legal advice, medical advice or in any way should be relied upon by anyone for any purpose. The Author does not provide any financial, investment, medical, legal or similar advice in this document or in its publications on any related Internet sites.

Contents

1 Introduction 3
2 The Target Genes and Housekeeping Genes 6
3 Data Extraction 8
4 Calculations 10
4.1 Classifiers 10
4.2 CCP Index 13
4.3 Combined Risk Score 16
5 Observations 18
6 References 20
7 Appendix A Genes 21

1 INTRODUCTION

There is always an interest in determining the prognostic value of tumors and hopefully staging treatment. There has been a recent flurry of interest in using cell cycle progression genes testing, a method of taking gene products from biopsy samples and then using them to ascertain the most likely progression of the tumor. CCP is a methodology proposed to do this. We take no position in this opinion paper regarding the efficacy of CCP as applied to PCa but we examine the original assertions in some detail. Conceptually it makes sense. It is as follows:

1. A handful of genes if over expressed, when combined with other metrics, can provide fairly accurate prognostic measures of PCa.

2. Selecting the genes can be accomplished in a variety of ways ranging from logical and clear pathway control genes such as PTEN to just a broad base sampling wherein the results have a statistically powerful predictive result.

3. Measuring the level of expression in some manner and from the measurements combine those in a reasonable fashion to determine a broad based metric.

4. Combining the gene expression metric with other variable to ascertain a stronger overall metric.

The CCP work to date has been focused somewhat on these objectives.

Let us now briefly update the work as detailed in the industry press. As indicated in a recent posting:[1]

Cuzick and his colleagues initially measured the levels of expression of a total of 31 genes involved in CCP. They used these data to develop a predefined CCP "score" and then they set out to evaluate the value of the CCP score in predicting risk for progressive disease in the men who had undergone an RP or risk of prostate cancer-specific mortality in the men who had been diagnosed by a TURP and managed by watchful waiting. The findings of this study can be summarized as follows:

Among patients in the two RP cohorts

1. *The CCP score could predict biochemical recurrence in univariate analysis (hazard ratio [HR] for a doubling in CCP = 1·89; p=5·6×10−9).*
2. *The CCP score could predict biochemical recurrence in the final multivariate analysis (HR =1·77; p=4·3×10−6).*
3. *The CCP score and the PSA level were the most important and the most clinically significant variables in the best predictive model (the final multivariate analysis).*

[1] http://prostatecancerinfolink.net/2011/02/09/is-ccp-testing-really-the-prognostic-tool-we-need/

Among patients in the TURP cohort

1. *The CCP score could predict time to death from prostate cancer in univariate analysis (HR = 2·92; p=6·1×10−22).*
2. *The CCP score could predict time of death from prostate cancer in the final multivariate analysis (HR = 2·57; p=8·2×10−11).*
3. *The CCP score was stronger than all other prognostic factors (although PSA levels added useful information).*

Thus there seems to be a strong belief in the use of CCP, especially when combined with other measures such as PSA.

The CCP test has been commercialized as Prolaris by Myriad. In a Medscape posting they state[2]:

The Prolaris test, which measures the activity of cell cycle progression (CCP) genes in prostate cancer biopsy samples, was evaluated for its ability to predict either death from prostate cancer or biochemical recurrence in 5 company-sponsored studies, Dr. Cuzick reported.

It was tested at the time of disease diagnosis in 2 conservatively managed cohorts from the United Kingdom (Lancet Oncol. 2011;12:245-255 and Br J Cancer. 2012;106:1095-1099), after radical prostatectomy in 2 cohorts from the United States (Lancet Oncol. 2011;12:245-255 and J Clin Oncol. 2013;31:1428-1434), and after external-beam radiation therapy (Freedland et al 2013, unpublished).

In the studies, formalin-fixed prostate tissue from men with prostate adenocarcinoma was analyzed. A CCP score was calculated by measuring the average RNA expression of 31 CCP genes normalized by the average expression of 15 housekeeping genes as quantitated with reverse-transcriptase polymerase chain reaction, explained Dr. Cuzick.

A hazard ratio was then calculated for every unit change in CCP score for the risk for either biochemical recurrence or death from prostate cancer.

"A unit change is essentially a doubling in the expression of these cell cycle genes," he explained.

On multivariate analysis — variables ranged in the different studies but all included Gleason score and prostate-specific antigen (PSA) level — the predictive value of the CCP score for either outcome was "dominant" and "hugely significant" (hazard ratio, 2.6; P < 1010), said Dr. Cuzick.

[2] http://www.medscape.com/viewarticle/805351

"PSA retained a fair amount of its predictive value, but the predictive value of the Gleason score "diminished" against the CCP score." he said. "Once you add the CCP score, there is little addition from the Gleason score, although there is some."

"Overall, the CCP score was a highly significant predictor of outcome in all of the studies," said Dr. Cuzick. "It was the dominant predictor in all but 1 of the studies in the multivariate analyses, and typically a unit change in the score was associated with a remarkably similar 2- to 3-fold increase in either death from prostate cancer or biochemical recurrence, indicating that this is a very robust predictor, and seems to work in a whole range of circumstances."

Thus there is some belief that CCP when combined with other metrics has strong prognostic value.

In this analysis we use CCP as both an end and a means to an end. CCP is one of many possible metrics to ascertain prognostic values. There is a wealth of them. We thus start with the selection of genes. The Appendix provides a description of all of them yet a more detailed pathway analysis is warranted but not included here not in the papers presented. We then examine classifiers for prognostic value. We first consider general issues and then apply them to the CCP approach. This is the area where we have the majority of our problems.

2 THE TARGET GENES AND HOUSEKEEPING GENES

The CCP sets of Target Genes and the Housekeeping Genes are depicted in the Table below. It is interesting to note that the Target Genes do not represent any of the usual suspects such as PTEN and cMYC. What is also of interest is what the pathway interactions are amongst the Target Genes. We shall focus on this at a later time (see the Appendix for details).

Target Gene	*Housekeeping Gene*
FOXM1	RPL38
CDC20	UBA52
CDKN3	PSMC1
CDC2	RPL4
KIF11	RPL37
KIAA0101	RPS29
NUSAP1	SLC25A3
CENPF	CLTC
ASPM	TXNL1
BUB1B	PSMA1
RRM2	RPL8
DLGAP5	MMADHC
BIRC5	RPL13A;LOC728658
KIF20A	PPP2CA
PLK1	MRFAP1
TOP2A	
TK1	
PBK	
ASF1B	
C18orf24	
RAD54L	
PTTG1	
CDCA3	
MCM10	
PRC1	
DTL	
CEP55	
RAD51	
CENPM	
CDCA8	
ORC6L	

The selection of the Target Genes is based purely upon the strength of its statistical predictive and prognostic value.

3 DATA EXTRACTION

Let us first examine how they obtained the data. We shall follow the text of the 2011 paper and then comment accordingly.

1. Extract RNA

Total RNA was extracted using either RNeasy FFPE or miRNeasy (Qiagen) as described by the manufacturer.

The miRNeasy kit became available after we had isolated RNA from about 1/3 of the RP cohort. We switched from RNeasy FFPE to miRNeasy because the new kit consistently generated better RNA yields. There was no difference in gene expression data between the kits.

2. Treat the RNA with enzyme to generate cDNA

Total RNA was treated with DNase I (Sigma) prior to cDNA synthesis.

We employed the High-capacity cDNA Archive Kit (Applied Biosystems) to convert total RNA into single strand cDNA as described by the manufacturer. Ideally, at least 200ng RNA was required for the RT reaction, but smaller input amounts were also successful. The quality of the RNA was not ideal, as is expected when isolating nucleic acids from old FFPE biopsies. Careful attention was given as to how to obtain a reliable score from this material in the development of this assay. RNA quality was determined via the amplifiability of the CCP and HK genes.

3. Collect the cDNA and confirm the generation of key entities.

In order to generate a CCP score, essentially all of the house-keeping genes and at least 21 CCP genes needed to amplify. We attempted to generate a CCP score from every sample. For some of the samples some genes failed to amplify indicating that the RNA quality was too poor to create a score. However, most samples (90% of the RP cohort and 85% of the TUPR cohort) generated CCP scores, and therefore, had adequate quality RNA.

4. Amplify the cDNA

Prior to measuring expression levels, the cDNA was pre-amplified with a pooled reaction containing TaqMan assays.

5. Pre amplify the cDNA prior to measuring in an array.

Pre-amplification reaction conditions were as follows: 14 cycles of 95°C for 15 seconds and 60°C for 4 minutes. The first cycle also included a ten minute incubation at 95°C.

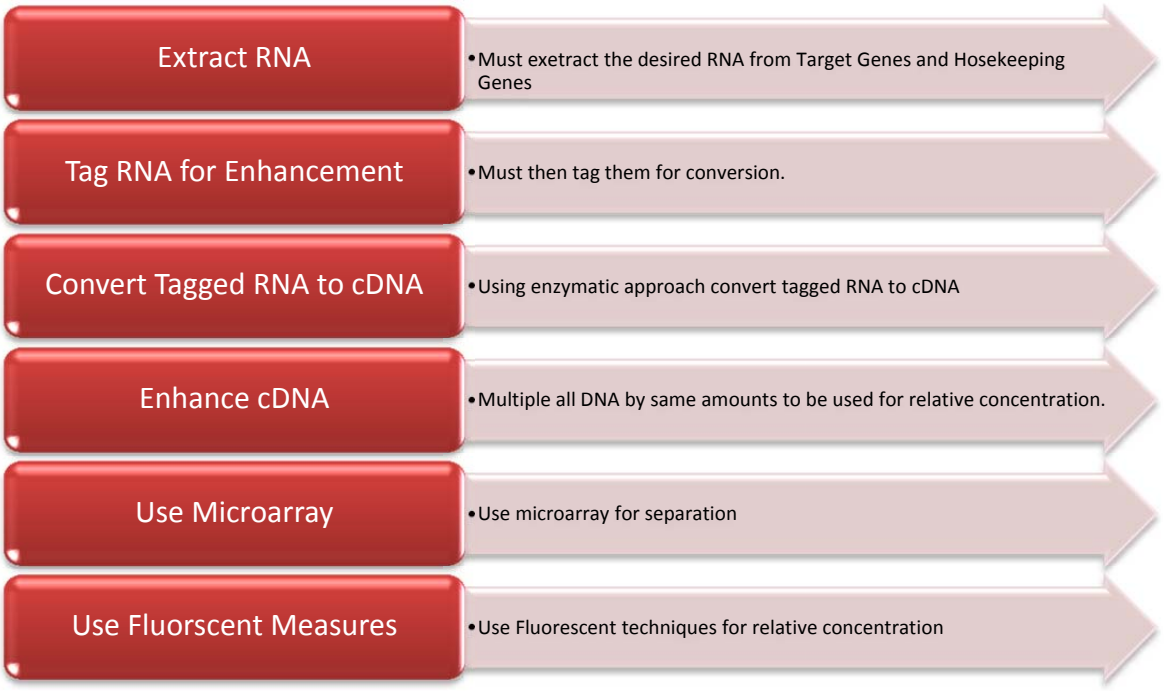
The amplification reaction was diluted 1:20 using the 1XTE buffer prior to loading on Taqman Low Density Arrays …to evaluate the amplified genes.

7. In arrays record levels of expression

Expression data were recorded as a C_T value, the PCR cycle at which the fluorescence intensity exceeded a predefined threshold. A total of 31 predefined CCP genes and 15 housekeeper genes were amplified on a single TLDA array.

Clearly there may be many sources of noise or error in this approach, especially in recording the level of fluorescent intensity.

We summarize the above process in the following graphic.



The problem is however that at each step we have the possibility of measurement bias or error. These become additive and can substantially alter the data results.

4 CALCULATIONS

In this section we consider the calculations needed to develop a reliable classifier. This is a long standing and classic problem. Simply stated:

"Assume you have N gene expression levels, G_i, and you desire to find some function $g(G_1,...,G_N)$ such that this function g divides the space created by the Gs into two regions, one with no disease progression and one with disease progression."

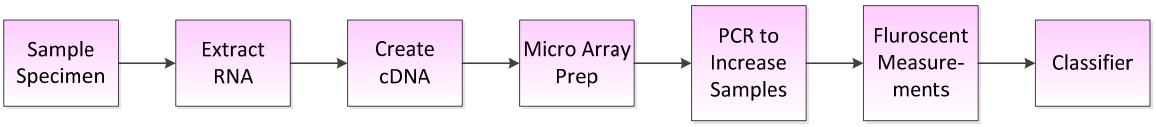
Alternatively we could ask for a function $f(G_1,...,G_N)$ such that the probability of disease progression, or an end point of death in a defined period, is f or some function derived therefrom. Namely we can determine:

$$P[Death\ in\ N\ months] = f(G_1,...,G_N)$$

This section will first discuss the general problem and then apply it to trying to interpret the CCP metric and then briefly look at a broader based metric.

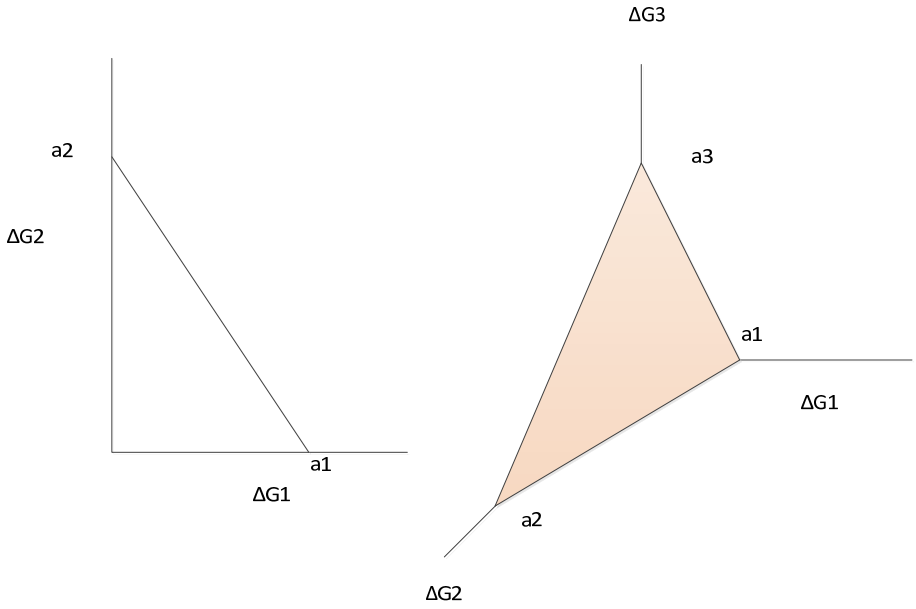
4.1 CLASSIFIERS

Let us begin with general classifiers. First let us review the process of collecting data. The general steps are below. We start with a specimen and we end up with N measurements of gene expression. In the CCP case we have some 31 genes we are examining and ascertaining their relative excess expression.



Now as we had posed the problem above we are seeking a classifier to determine a function f or g as above which would either bifurcate the space of N genes or a function f from which we could ascertain survival based upon the N gene expression measurements.

Now from classic classifier analysis we can develop the two metrics; a simple bifurcating classifier and a probability estimator. The simple classifier generates a separation point, a line or plane as shown below, for which being below is benign and being above is problematic. This is akin to the simple PSA test of being above or below 4.0. However we all know that this has its problems.



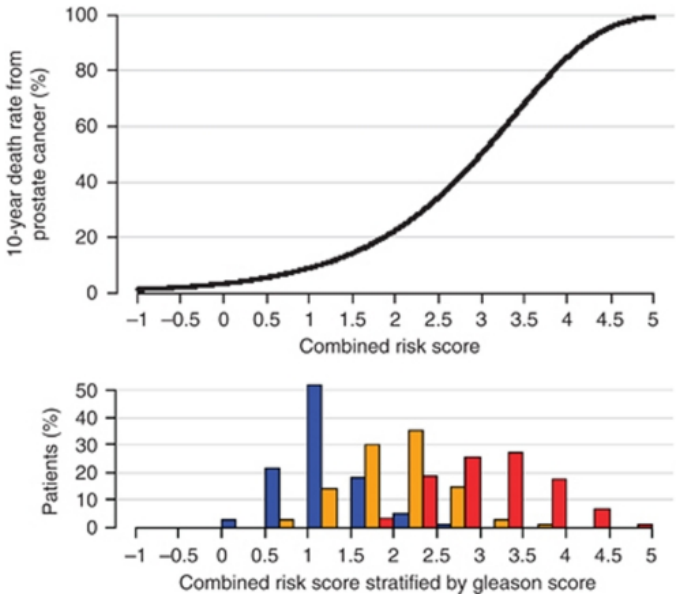
Now the probability estimator approach has more merit. The problem is how one determines the metric to do this. Frequently we generate a simple metric such as the following:

$$M = \sum_{i=1}^{N} a_i \Delta G(i)$$

Thus we have the excess gene expression measures and now we need to determine the "a" weights to provide a best M metric.

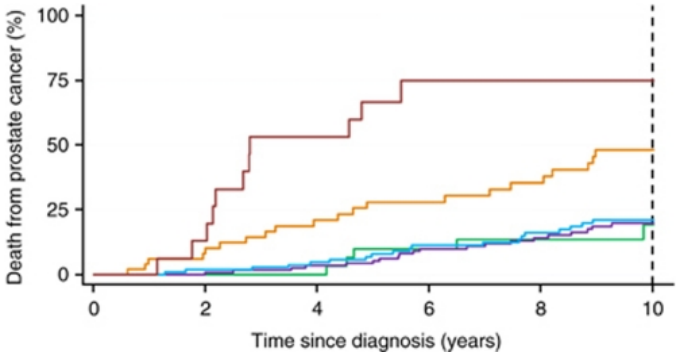
The Figure below is from Cuzick et al and demonstrates their prognostic curve[3].

[3] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3304411/#!po=3.84615



Note the combined risk score is the metric they have proposed and which we shall consider later in this section. Namely they take the 31 gene expression levels and combine them to a single metric and from that can ascertain the prognostic value.

The actual data from Cuzick et al is shown below in Kaplan Meir form:



Where they state: "*Different categories of CCP score are shown by different coloured lines: red, CCP score>3, orange, 2<CCP score <3; blue, 1<CCP score <2; purple, 0<CCP score <1; green, CCP score <0.*"

Thus there may be some validity in the approach for prognostic purposes. Clearly a high value indicates a significant chance for mortality, one assumes directly related to this disease.

Now if we examine the above survival graph we can pose the following problem. This perhaps what the authors may be attempting to do.

Let us assume we have a set of data; $\{G_1,...,G_N\}$ so that we can look at an N dimension space from which our data arises. We could further assume, without basis of course, that these measurements are independent variables. Then we could ask to estimate:

$$P\left[\, Death\ at\ Month\ k \,\middle|\, \{G_1,...,G_N\} \right]$$

This means that we have conditioned the probability on an N dimensional data set in $\mathbb{R}^N$. However this, as we shall see is not what the author does. He assumes that there is a mapping from $\mathbb{R}^N$ to $\mathbb{R}^1$.Namely some g exists such that:

$$g = g(G_1,...,G_N)$$

and thus we obtain:

$$P\left[\, Death\ at\ Month\ k \,\middle|\, g \right]$$

The mapping from the N dimensional space to the one dimensional space is problematic at best. This g is the CRs, the combined risk score or element of it. This is a highly complex issue which unfortunately in my opinion one finds impossible to understand from the paper.

4.2 CCP INDEX

Let us now examine the CCP index calculation in some detail. We use the Cuzick et al 2011 paper as the source. The subsequent papers refer back to this and thus we rely upon what little is presented here. The approach we take herein is to use what the original paper stated and then line by line establish a mathematical model and where concerns or ambiguities we point them out for subsequent resolution. In our opinion the presentation of the quantitative model is seriously flawed in terms of its explanation and we shall show the basis of our opinion below.

Let us commence with how it is explained in the original paper:

Expression data were recorded as a C_T value, the PCR cycle at which the fluorescence intensity exceeded a predefined threshold. A total of 31 predefined CCP genes and 15 housekeeper genes were amplified on a single TLDA array.

Now this I assume means that we have the following data:

$$C_T^{TG}(k); k = 1,31$$
$$C_T^{HK}(j); j = 1,15$$

Where we have Target Genes and Housekeeping genes. Namely we have relative expression levels for 31+15 or 46 genes. There is no discussion, as far as I could discern, as to the possible values for these expression values or the standardized process for collecting them.

The CCP score for each individual was calculated as follows:

*For **each** of **three replicates** of each of the **31 CCP genes**,*

C_T *values were normalised by*

subtracting the average of up to 15 non-failed housekeeper (HK) genes (centered using a predefined value) to yield ΔC_T^* .

Now one could assume from what is stated above from the paper that the following mathematical process is performed. Take 3 copies of a specific TG, say k=n, and calculate the following:

Define :

$$E[C_T^{HK}] = \frac{\sum_{j=1}^{J} C_T^{HK}(j)}{J} = \overline{C_T^{HK}}$$

then

$$\Delta C_T^{TG}(k) = C_T^{TG}(k) - C_T^{HK}$$

Now this number is for a single sample of a single Target Gene. Based upon what we have read we have three such samples and 31 such Target Genes. Thus we have 93 values calculated. One would think.

Then, a predefined baseline value ΔC_T^* *was subtracted from to create a quantity labelled* $\Delta\Delta C_T^*$

One must assume that some specific number, the so-called "predefined baseline value" is then subtracted, the number whose basis is unknown and whose value is unknown. This of course, if properly interpreted, makes the analysis unrepeatable. Namely we have:

$$\Delta\Delta C_T^* = \Delta C_T^* - BL$$

Again it is fair to say we have a similar 93 such values.

But let us continue:

This was then converted to a quantity proportional to copy number, calculated as $2^{-\Delta\Delta C_T^*}$

One cannot, in my opinion, specifically understand how the authors want to use the copy number reference. Now again one must assume we have some 93 such quantities. The larger the exponent the smaller the value, this is a simple observation. One again wonders about that arbitrary offset that was subtracted, namely the BL value. That is some form of bias.

For missing $2^{-\Delta\Delta C_T^*}$ *values due to low expression,* $2^{-\Delta\Delta C_T^*}$ *was set equal to 0.*

That clearly follows. But let us continue to assume we have all the TG and HK genes.

The mean was calculated for each CCP gene as the mean $2^{-\Delta\Delta C_T^*}$ *of the qualifying replicates, i.e. those with expression of at least 13 HK genes, which was then averaged over the qualifying CCP genes.*

Now one needs to interpret a bit this statement. One assumes that we have 3 of the TG CCP genes and that we then calculate the mean. The mean of what set of numbers? It appears that we calculate the mean of $2^{-\Delta\Delta C_T^*}$ for the three genes. Thus we assume we calculate for each TG:

$$\overline{2^{-\Delta\Delta C_T^*}} = \frac{1}{3}\sum_{i=1}^{3} 2^{-\Delta\Delta C_T^*}(i)$$

A CCP gene was considered failed if more than one replicate did not qualify, or if two replicates qualified and one of them had $2^{-\Delta\Delta C_T^*}$ *equal to zero, or if the standard deviation between the three replicate values* $\Delta\Delta C_T^*$ *exceeded 0.5.*

The above is some form of exclusion factor. Again as with all that is presented there is no logic or basis. We will continue and assume that no such exclusion has occurred.

Finally, this was converted back to the CCP score by taking a base 2 logarithm.

Now somehow it is argued that we take this average and return it to a number. The logic for doing this is missing. First look at the wording. It is converted back to a score which was never in existence in the first place.

The goal is to get some CCP score. Let us consider two options.

Option 1: Average of Scores

We assume we follow the above and then take the log of each average scored and then average that. This is as follows:

Recall

$$\overline{2^{-\Delta\Delta C_T^*}} = \frac{1}{3}\sum_{i=1}^{3} 2^{-\Delta\Delta C_T^*}(i)$$

then

$$CCPS(i) = log_2 \overline{2^{-\Delta\Delta C_T^*}}$$

and

$$CCPS = \frac{1}{N}\sum_{i=1}^{N} CCPS(i)$$

Option 2: Score of Averages

This assumes we average the values first and then taken the log.

Then

$$\overline{2^{-\Delta\Delta C_T^*}}(j)$$

exists

and

$$\overline{2^{-\Delta\Delta C_T^*}} = \frac{1}{N}\sum_{j=1}^{N} \overline{2^{-\Delta\Delta C_T^*}}(j)$$

and

$$CCPS = log_2 \overline{2^{-\Delta\Delta C_T^*}}$$

These are two different functions yielding two different numbers. The paper has great ambiguity on this point. One finds it highly problematic in interpreting this algorithm.

CCP scores with the number of failing CCP genes in excess of 9 out of the 31, or a high standard deviation between scores calculated from the three replicates, were rejected and excluded from the analysis.

The interassay variability has been established in our laboratory and the standard deviation of the CCP score for experimental replicates is 0.1.

Now one asks what do we do with this number and how large can it be, or how small. For example, if we have an excess expression of say 2, then the 0.25 and the log to the base 2 of that is -2. What does that do for us? In my opinion, there is a great deal of confusion here.

4.3 COMBINED RISK SCORE

The best predictor based on these variables was suggested by Cusick et al as follows:

*Combined Risk Score = 0·55*CCP + 0·81*log(1+PSA) + 0·28*T-stage + 0·64*Margins { + 0·30*(Gleason = 7) + 0·99*(Gleason > 7)}*

One can restate this also as follows in a clearer form:

$$CRS = 0.55CCPS + 0.81log(1 + PSA) + 0.28TSS + 0.64\left(0.30 * G7 + 0.99G8\right)$$

where

$G7 = 0,1$

$G8 = 1,0$

As with so many other issues in this paper there is no base to the log.

5 OBSERVATIONS

his area of investigation is of interest but it in my opinion raises more questions than posing answers. First is the issue of the calculation itself and its reproducibility. Second is the issue of the substantial noise inherent in the capture of the data.

1. Pathway Implications: Is this just another list of Genes?

The first concern is the fact that we know a great deal about ligands, receptors, pathway elements, and transcription factors. Why, one wonders, do we seem to totally neglect that source of information.

2. Noise Factors: The number of genes and the uncertainties in measurements raise serious concerns as to stability of outcomes.

Noise can be a severe detractor from the usefulness of the measurement. There are many sources of such noise especially in measuring the fluorescent intensity. One wonders how they factor into the analysis. Many others sources are also present from the PCR process and copy numbers to the very sampling and tissue integrity factors.

3. Severity of Prognosis and Basis: For a measurement which is predicting patient death one would expect total transparency.

The CCP discriminant argues for the most severe prognostication. Namely it dictates death based upon specific discriminant values. However as we have just noted, measurement noise can and most likely will provide significant uncertainty in the “true” value of the metric.
4. Flaws in the Calculation Process: Independent of the lack of apparent transparency, there appear in my opinion to be multiple points of confusion in the exposition of the methodology.

In our opinion, there are multiple deficiencies in the presentation of the desired calculation of the metric proposed which make it impossible to reproduce it. We detail them in our White Paper.

5. Discriminants, Classifiers, Probability Estimators: What are they really trying to do?

The classic question when one has N independent genes and when one can measure relative expression is how does one take that data and determine a discriminant function. All too often the intent is to determine a linear one dimensional discriminant. At the other extreme is a multidimensional non-linear discriminant. This is always the critical issue that has been a part of classifiers since the early 1950s. In the case considered herein there is little if any description of or justification of the method employed. One could assume that the authors are trying to obtain an estimate of the following:

P[Death in M months]=g(G1,...GN))

where Gk is the level of expression of one of the 31 genes. One would immediately ask; why and how? In fact we would be asked to estimate a Bayesian measure:

P[Death in M months|G1,...GN]

which states that we want the conditional probability. We know how to do this for systems but this appears at best to be some observational measure. This in my opinion is one of the weak points.

6. Causal Genes, where are they?

One of the major concerns is that one genes expression is caused by another gene. In this case of 31 genes there may be some causality and thus this may often skew results.

7. Which Cell?

One of the classic problems is measuring the right cell. Do we want the stem cell, if so how are they found. Do we want metastatic cells, then from where do we get them. Do we want just local biopsy cells, if so perhaps they under-express the facts.

8. Why this when we have so many others?

We have PSA, albeit with issues, we have SNPs, we have ligands, receptors, pathway elements, transcription factors, miRNAs and the list goes on. What is truly causal?

Basically this approach has possible merit. The problem, in my opinion, is the lack of transparency in the description of the test metric. Also the inherent noisy data is a concern in my opinion. Moreover one wonders why so much Press.

6 REFERENCES

1. Cooperberg, M., et al, Validation of a Cell-Cycle Progression Gene Panel to Improve! Risk Stratification in a Contemporary Prostatectomy Cohort! https://s3.amazonaws.com/myriad-library/Prolaris/UCSF+ASCO+GU.pdf

2. Cooperberg, M., et al, Validation of a Cell-Cycle Progression Gene Panel to Improve Risk Stratification in a Contemporary Prostatectomy Cohort, JOURNAL OF CLINICAL ONCOLOGY, 2012.

3. Cuzick J., et al, Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort, British Journal of Cancer (2012) 106, 1095 – 1099.

4. Cuzick, J., et al, Prognostic value of an RNA expression signature derived from cell cycle proliferation genes for recurrence and death from prostate cancer: A retrospective study in two cohorts, Lancet Oncol. 2011 March; 12(3): 245–255.

5. Duda, R., et al, Pattern Classification, Wiley (New York) 2001.

6. McGarty, T., Prostate Cancer Genomics, Draft 2, 2013, http://www.telmarc.com/Documents/Books/Prostate%20Cancer%20Systems%20Approach%2003.pdf

7. Theodoridis, S., K., Koutroumbas, Pattern Recognition, AP (New York) 2009.

7 APPENDIX A GENES

The following is a detailed Table of the Target Genes. They are from NCBI[4].

[4] http://www.ncbi.nlm.nih.gov/

Target Gene	Description
FOXM1	The protein encoded by this gene is a transcriptional activator involved in cell proliferation. The encoded protein is phosphorylated in M phase and regulates the expression of several cell cycle genes, such as cyclin B1 and cyclin D1.
CDC20	CDC20 appears to act as a regulatory protein interacting with several other proteins at multiple points in the cell cycle. It is required for two microtubule-dependent processes, nuclear movement prior to anaphase and chromosome separation.
CDKN3	The protein encoded by this gene belongs to the dual specificity protein phosphatase family. It was identified as a cyclin-dependent kinase inhibitor, and has been shown to interact with, and dephosphorylate CDK2 kinase, thus prevent the activation of CDK2 kinase. This gene was reported to be deleted, mutated, or overexpressed in several kinds of cancers. Alternatively spliced transcript variants encoding different isoforms have been found for this gene.
CDC2	The protein encoded by this gene is a member of the Ser/Thr protein kinase family. This protein is a catalytic subunit of the highly conserved protein kinase complex known as M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of eukaryotic cell cycle. Mitotic cyclins stably associate with this protein and function as regulatory subunits. The kinase activity of this protein is controlled by cyclin accumulation and destruction through the cell cycle. The phosphorylation and dephosphorylation of this protein also play important regulatory roles in cell cycle control. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. Also CDK1
KIF11	This gene encodes a motor protein that belongs to the kinesin-like protein family. Members of this protein family are known to be involved in various kinds of spindle dynamics. The function of this gene product includes chromosome positioning, centrosome separation and establishing a bipolar spindle during cell mitosis.
KIAA0101	HCV NS5A-transactivated protein 9; PCNA-associated factor; PCNA-associated factor of 15 kDa; hepatitis C virus NS5A-transactivated protein 9; overexpressed in anaplastic thyroid carcinoma 1
NUSAP1	NUSAP1 is a nucleolar-spindle-associated protein that plays a role in spindle microtubule organization
CENPF	This gene encodes a protein that associates with the centromere-kinetochore complex. The protein is a component of the nuclear matrix during the G2 phase of interphase. In late G2 the protein associates with the kinetochore and maintains this association through early anaphase. It localizes to the spindle midzone and the intracellular bridge in late anaphase and telophase, respectively, and is thought to be subsequently degraded. The localization of this protein suggests that it may play a role in chromosome segregation during mitotis. It is thought to form either a homodimer or heterodimer. Autoantibodies against this protein have been found in patients with cancer

	or graft versus host disease.
ASPM	This gene is the human ortholog of the Drosophila melanogaster 'abnormal spindle' gene (asp), which is essential for normal mitotic spindle function in embryonic neuroblasts. Studies in mouse also suggest a role of this gene in mitotic spindle regulation, with a preferential role in regulating neurogenesis. Mutations in this gene are associated with microcephaly primary type 5. Multiple transcript variants encoding different isoforms have been found for this gene.
BUB1B	This gene encodes a kinase involved in spindle checkpoint function. The protein has been localized to the kinetochore and plays a role in the inhibition of the anaphase-promoting complex/cyclosome (APC/C), delaying the onset of anaphase and ensuring proper chromosome segregation. Impaired spindle checkpoint function has been found in many forms of cancer.
RRM2	This gene encodes one of two non-identical subunits for ribonucleotide reductase. This reductase catalyzes the formation of deoxyribonucleotides from ribonucleotides. Synthesis of the encoded protein (M2) is regulated in a cell-cycle dependent fashion. Transcription from this gene can initiate from alternative promoters, which results in two isoforms that differ in the lengths of their N-termini. Related pseudogenes have been identified on chromosomes 1 and X.
DLGAP5	discs, large (Drosophila) homolog-associated protein 5
BIRC5	This gene is a member of the inhibitor of apoptosis (IAP) gene family, which encodes negative regulatory proteins that prevent apoptotic cell death. IAP family members usually contain multiple baculovirus IAP repeat (BIR) domains, but this gene encodes proteins with only a single BIR domain. The encoded proteins also lack a C-terminus RING finger domain. Gene expression is high during fetal development and in most tumors, yet low in adult tissues. Alternatively spliced transcript variants encoding distinct isoforms have been found for this gene.
KIF20A	kinesin family member 20A
PLK1	polo-like kinase 1
TOP2A	his gene encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription. This nuclear enzyme is involved in processes such as chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication. It catalyzes the transient breaking and rejoining of two strands of duplex DNA which allows the strands to pass through one another, thus altering the topology of DNA. Two forms of this enzyme exist as likely products of a gene duplication event. The gene encoding this form, alpha, is localized to chromosome 17 and the beta gene is localized to chromosome 3. The gene encoding this enzyme functions as the target for several anticancer agents and a variety of mutations in this gene have been associated with the development of drug resistance. Reduced activity of this enzyme may also play a role in ataxia-

	telangiectasia.
TK1	thymidine kinase 1, soluble
PBK	This gene encodes a serine/threonine kinase related to the dual specific mitogen-activated protein kinase kinase (MAPKK) family. Evidence suggests that mitotic phosphorylation is required for its catalytic activity. This mitotic kinase may be involved in the activation of lymphoid cells and support testicular functions, with a suggested role in the process of spermatogenesis.
ASF1B	This gene encodes a member of the H3/H4 family of histone chaperone proteins and is similar to the anti-silencing function-1 gene in yeast. The encoded protein is the substrate of the tousled-like kinase family of cell cycle-regulated kinases, and may play a key role in modulating the nucleosome structure of chromatin by ensuring a constant supply of histones at sites of nucleosome assembly.
C18orf24	Also SKA1 spindle and kinetochore associated complex subunit 1
RAD54L	The protein encoded by this gene belongs to the DEAD-like helicase superfamily, and shares similarity with Saccharomyces cerevisiae Rad54, a protein known to be involved in the homologous recombination and repair of DNA. This protein has been shown to play a role in homologous recombination related repair of DNA double-strand breaks. The binding of this protein to double-strand DNA induces a DNA topological change, which is thought to facilitate homologous DNA paring, and stimulate DNA recombination. Alternative splicing results in multiple transcript variants encoding the same protein
PTTG1	The encoded protein is a homolog of yeast securin proteins, which prevent separins from promoting sister chromatid separation. It is an anaphase-promoting complex (APC) substrate that associates with a separin until activation of the APC. The gene product has transforming activity in vitro and tumorigenic activity in vivo, and the gene is highly expressed in various tumors. The gene product contains 2 PXXP motifs, which are required for its transforming and tumorigenic activities, as well as for its stimulation of basic fibroblast growth factor expression. It also contains a destruction box (D box) that is required for its degradation by the APC. The acidic C-terminal region of the encoded protein can act as a transactivation domain. The gene product is mainly a cytosolic protein, although it partially localizes in the nucleus.
CDCA3	cell division cycle associated 3
MCM10	The protein encoded by this gene is one of the highly conserved mini-chromosome maintenance proteins (MCM) that are involved in the initiation of eukaryotic genome replication. The hexameric protein complex formed by MCM proteins is a key component of the pre-replication complex (pre-RC) and it may be involved in the formation of replication forks and in the recruitment of other DNA replication related proteins. This protein can interact with MCM2 and MCM6, as well as with the origin recognition protein ORC2. It is regulated by proteolysis and phosphorylation in a cell

	cycle-dependent manner. Studies of a similar protein in Xenopus suggest that the chromatin binding of this protein at the onset of DNA replication is after pre-RC assembly and before origin unwinding. Alternatively spliced transcript variants encoding distinct isoforms have been identified.
PRC1	This gene encodes a protein that is involved in cytokinesis. The protein is present at high levels during the S and G2/M phases of mitosis but its levels drop dramatically when the cell exits mitosis and enters the G1 phase. It is located in the nucleus during interphase, becomes associated with mitotic spindles in a highly dynamic manner during mitosis, and localizes to the cell mid-body during cytokinesis. This protein has been shown to be a substrate of several cyclin-dependent kinases (CDKs). It is necessary for polarizing parallel microtubules and concentrating the factors responsible for contractile ring assembly. Alternative splicing results in multiple transcript variants.
DTL	denticleless E3 ubiquitin protein ligase homolog
CEP55	centrosomal protein 55kDa
RAD51	The protein encoded by this gene is a member of the RAD51 protein family. RAD51 family members are highly similar to bacterial RecA and Saccharomyces cerevisiae Rad51, and are known to be involved in the homologous recombination and repair of DNA. This protein can interact with the ssDNA-binding protein RPA and RAD52, and it is thought to play roles in homologous pairing and strand transfer of DNA. This protein is also found to interact with BRCA1 and BRCA2, which may be important for the cellular response to DNA damage. BRCA2 is shown to regulate both the intracellular localization and DNA-binding ability of this protein. Loss of these controls following BRCA2 inactivation may be a key event leading to genomic instability and tumorigenesis. Multiple transcript variants encoding different isoforms have been found for this gene.
CENPM	The centromere is a specialized chromatin domain, present throughout the cell cycle that acts as a platform on which the transient assembly of the kinetochore occurs during mitosis. All active centromeres are characterized by the presence of long arrays of nucleosomes in which CENPA (MIM 117139) replaces histone H3 (see MIM 601128). CENPM is an additional factor required for centromere assembly
CDCA8	This gene encodes a component of the chromosomal passenger complex. This complex is an essential regulator of mitosis and cell division. This protein is cell-cycle regulated and is required for chromatin-induced microtubule stabilization and spindle formation. Alternate splicing results in multiple transcript variants. Pseudgenes of this gene are found on chromosomes 7, 8 and 16. [
ORC6L	The origin recognition complex (ORC) is a highly conserved six subunit protein complex essential for the initiation of the DNA replication in eukaryotic cells. Studies in yeast demonstrated that ORC binds specifically to origins of replication and serves as a platform for the assembly of additional initiation factors such as Cdc6 and Mcm proteins. The protein

	encoded by this gene is a subunit of the ORC complex. Gene silencing studies with small interfering RNA demonstrated that this protein plays an essential role in coordinating chromosome replication and segregation with cytokinesis

Housekeeping Gene	*Description*
RPL38	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L38E family of ribosomal proteins. It is located in the cytoplasm. Alternative splice variants have been identified, both encoding the same protein. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome, including one located in the promoter region of the type 1 angiotensin II receptor gene.
UBA52	Ubiquitin is a highly conserved nuclear and cytoplasmic protein that has a major role in targeting cellular proteins for degradation by the 26S proteosome. It is also involved in the maintenance of chromatin structure, the regulation of gene expression, and the stress response. Ubiquitin is synthesized as a precursor protein consisting of either polyubiquitin chains or a single ubiquitin moiety fused to an unrelated protein. This gene encodes a fusion protein consisting of ubiquitin at the N terminus and ribosomal protein L40 at the C terminus, a C-terminal extension protein (CEP). Multiple processed pseudogenes derived from this gene are present in the genome.
PSMC1	The 26S proteasome is a multicatalytic proteinase complex with a highly ordered structure composed of 2 complexes, a 20S core and a 19S regulator. The 20S core is composed of 4 rings of 28 non-identical subunits; 2 rings are composed of 7 alpha subunits and 2 rings are composed of 7 beta subunits. The 19S regulator is composed of a base, which contains 6 ATPase subunits and 2 non-ATPase subunits, and a lid, which contains up to 10 non-ATPase subunits. Proteasomes are distributed throughout eukaryotic cells at a high concentration and cleave peptides in an ATP/ubiquitin-dependent process in a non-lysosomal pathway. An essential function of a modified proteasome, the immunoproteasome, is the processing of class I MHC peptides. This gene encodes one of the ATPase subunits, a member of the triple-A family of ATPases which have a chaperone-like activity. This subunit and a 20S core alpha subunit interact specifically with the hepatitis B virus X protein, a protein critical to viral replication. This subunit also interacts with the adenovirus E1A protein and this interaction alters the activity of the proteasome. Finally, this subunit interacts with ataxin-7, suggesting a role for the proteasome in the development of spinocerebellar ataxia type 7, a progressive neurodegenerative disorder.
RPL4	Ribosomes, the organelles that catalyze protein synthesis, consist of

	a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L4E family of ribosomal proteins. It is located in the cytoplasm. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome.
RPL37	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L37E family of ribosomal proteins. It is located in the cytoplasm. The protein contains a C2C2-type zinc finger-like motif. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome.
RPS29	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 40S subunit and a member of the S14P family of ribosomal proteins. The protein, which contains a C2-C2 zinc finger-like domain that can bind to zinc, can enhance the tumor suppressor activity of Ras-related protein 1A (KREV1). It is located in the cytoplasm. Variable expression of this gene in colorectal cancers compared to adjacent normal tissues has been observed, although no correlation between the level of expression and the severity of the disease has been found. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. [
SLC25A3	The protein encoded by this gene catalyzes the transport of phosphate into the mitochondrial matrix, either by proton cotransport or in exchange for hydroxyl ions. The protein contains three related segments arranged in tandem which are related to those found in other characterized members of the mitochondrial carrier family. Both the N-terminal and C-terminal regions of this protein protrude toward the cytosol. Multiple alternatively spliced transcript variants have been isolated.
CLTC	Clathrin is a major protein component of the cytoplasmic face of intracellular organelles, called coated vesicles and coated pits. These specialized organelles are involved in the intracellular trafficking of receptors and endocytosis of a variety of macromolecules. The basic subunit of the clathrin coat is composed

	of three heavy chains and three light chains.
TXNL1	thioredoxin-like 1
PSMA1	The proteasome is a multicatalytic proteinase complex with a highly ordered ring-shaped 20S core structure. The core structure is composed of 4 rings of 28 non-identical subunits; 2 rings are composed of 7 alpha subunits and 2 rings are composed of 7 beta subunits. Proteasomes are distributed throughout eukaryotic cells at a high concentration and cleave peptides in an ATP/ubiquitin-dependent process in a non-lysosomal pathway. An essential function of a modified proteasome, the immunoproteasome, is the processing of class I MHC peptides. This gene encodes a member of the peptidase T1A family, that is a 20S core alpha subunit. Alternative splicing results in multiple transcript variants encoding distinct isoforms
RPL8	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L2P family of ribosomal proteins. It is located in the cytoplasm. In rat, the protein associates with the 5.8S rRNA, very likely participates in the binding of aminoacyl-tRNA, and is a constituent of the elongation factor 2-binding site at the ribosomal subunit interface. Alternatively spliced transcript variants encoding the same protein exist. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed through the genome.
MMADHC	This gene encodes a mitochondrial protein that is involved in an early step of vitamin B12 metabolism. Vitamin B12 (cobalamin) is essential for normal development and survival in humans. Mutations in this gene cause methylmalonic aciduria and homocystinuria type cblD (MMADHC), a disorder of cobalamin metabolism that is characterized by decreased levels of the coenzymes adenosylcobalamin and methylcobalamin. Pseudogenes have been identified on chromosomes 11 and X.
RPL13A;LOC728658	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins. This gene encodes a member of the L13P family of ribosomal proteins that is a component of the 60S subunit. The encoded protein also plays a role in the repression of inflammatory genes as a component of the IFN-gamma-activated inhibitor of translation (GAIT) complex. This gene is co-transcribed with the small nucleolar RNA genes U32, U33, U34, and U35, which are located in the second, fourth, fifth, and sixth introns,

	respectively. As is typical for genes encoding ribosomal proteins, there are multiple processed pseudogenes of this gene dispersed throughout the genome. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene.
PPP2CA	This gene encodes the phosphatase 2A catalytic subunit. Protein phosphatase 2A is one of the four major Ser/Thr phosphatases, and it is implicated in the negative control of cell growth and division. It consists of a common heteromeric core enzyme, which is composed of a catalytic subunit and a constant regulatory subunit, that associates with a variety of regulatory subunits. This gene encodes an alpha isoform of the catalytic subunit.
MRFAP1	This gene encodes an intracellular protein that interacts with members of the MORF4/MRG (mortality factor on chromosome 4/MORF4 related gene) family and the tumor suppressor Rb (retinoblastoma protein.) The protein may play a role in senescence, cell growth and immortalization. Alternative splicing results in multiple transcript variants.