Published in final edited form as:

J Immunol Methods. 2012 February 28; 376(1-2): 108–112. doi:10.1016/j.jim.2011.12.003.

Pitfalls in retrospective analyses of biomarkers: A case study with metastatic melanoma patients

Douglas M. Potter^{a,b,*}, Lisa H. Butterfield^{c,d,e}, Sherrie J. Divito^e, Cindy A. Sander^c, and John M. Kirkwood^c

Lisa H. Butterfield: butterfieldl@upmc.edu; Sherrie J. Divito: sdivito@partners.org; Cindy A. Sander: sandca@upmc.edu; John M. Kirkwood: kirkwoodjm@upmc.edu

^aBiostatistics Facility, University of Pittsburgh Cancer Institute, 201 North Craig Street, Suite 325, Pittsburgh, PA 15213, United States

^bDepartment of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 201 North Craig Street, Suite 325, Pittsburgh, PA 15213, United States

^cDepartments of Medicine, Dermatology and Translational Science, University of Pittsburgh Cancer Institute and University of Pittsburgh School of Medicine, 5117 Centre Avenue, Pittsburgh, PA 15213, United States

^dDepartment of Surgery, University of Pittsburgh Cancer Institute and University of Pittsburgh School of Medicine, 5117 Centre Avenue, Pittsburgh, PA 15213, United States

^eDepartment of Immunology, University of Pittsburgh Cancer Institute and University of Pittsburgh School of Medicine, 5117 Centre Avenue, Pittsburgh, PA 15213, United States

Abstract

Background—Reliable prognostic biomarkers of survival and response to treatment are clearly important in oncology, and many studies have been carried out with the objective of identifying new prognostic biomarkers. Retrospective analysis of blood banked from patients is a frequently used paradigm for these studies. We describe a new study of the association of serum biomarker level with overall survival in melanoma patients, and the problems encountered in carrying it out.

Methods—Blood samples from 56 patients with stage IV metastatic melanoma were drawn prior to initiation of any treatment for their disease. Sera from the samples were stored for up to 94 months at -80°C , and were subsequently thawed at the same time and tested by multiplex Luminex assay for 30 analytes (cytokines, chemokines and growth factors). Cox regression analysis was used to assess the association between these analytes and time-to-death.

Results—Of the 30 analytes, 17 were associated with survival, most strongly so, and in all cases, a higher analyte level was associated with increased survival. In addition, the correlations of the levels of all possible pairs of analytes were all positive and in almost all cases highly significant. However, these results are artifacts that arise from the combination of two peculiarities of the data: the apparent decrease in analyte level with storage time, and the uniformly shorter storage times of the samples from censored patients than the storage times of the samples from patients who died.

© 2011 Elsevier B.V. All rights reserved.

Corresponding author: Tel.: +1 412 383 1580; fax: +1 412 383 1535, dmp71@pitt.edu (D.M. Potter).

Authors' contributions

LHB participated in the design of the study, oversaw the assays and helped to draft the manuscript. SJD participated in the design of the study and helped to draft the manuscript. JMK conceived of the design. DMP performed the data analyses and was primarily responsible for drafting the manuscript. CAS was responsible for sample selection. All authors read and approved the final manuscript.

Conclusions—All retrospective studies can have hidden biases, and thus investigators should not claim new findings before examining the data in detail with the goal of determining whether the findings could be spurious. There were several suspicious findings in our initial analyses: too many analytes found significant, too many very small p-values, a uniformly positive association of analyte level with survival, and a uniformly positive correlation between analyte levels. We were convinced that these findings must be artifacts, and further analyses showed that the findings could be explained by an apparent decrease of analyte level storage time.

Keywords

Biomarker; Bias; Retrospective; Metastatic melanoma; Survival

1. Introduction

The incidence of melanoma continues to rise worldwide (American Cancer Society), and despite the high cure rate of melanoma that is detected and treated at an early stage (Kim et al., 2002), the prognosis of metastatic melanoma remains quite poor, with an average 5-year survival rate of 11% and median survival time between 6 and 9 months. Compounding this problem, there has been no evidence of improvement in the survival of melanoma with medical interventions (Korn et al., 2008; Balch, 1997; Balch et al., 2001; Thompson et al., 2005). Nonetheless, a small fraction of patients with stage IV melanoma do survive more than 5 years, often regardless of treatment. Clearly, it would be useful to identify reliable prognostic indicators of survival and response to treatment, and there has been considerable research aimed at achieving this goal. We describe a study of the association of serum biomarker level with overall survival, and the problems encountered in carrying it out. The biomarkers consisted of a panel of 30 cytokines, chemokines and growth factors present in blood.

2. Methods

2.1. Patients and sera

Sera from 56 patients were obtained from a serum bank, for which samples were obtained using the blood collection protocol UPCI # 96–099 IRB#970186. Patients were seen at the outpatient Melanoma Center of the University of Pittsburgh Cancer Institute (UPCI) between 2001 and 2007. Each patient gave written informed consent for the collection of peripheral blood and the processing of serum and blood lymphocytes. Blood was collected after initial evaluation and confirmation of the diagnosis of stage IV metastatic melanoma, but before initiation of any systemic treatment for metastatic disease. Venous blood was drawn into red top serum tubes (no anti-coagulant) (BD Biosciences, Franklin Lakes, NJ), clotted at room temperature for at least 30 min, and then centrifuged at 2500 *g* for 10 min. Aliquots of patient sera were then stored in a freezer at –80 °C, typically within 4 h of the blood draw, but always within 16 h. The freezers were monitored 24 h a day for temperature changes, and technical personnel were on call in case of equipment malfunction. All procedures were carried out in the University of Pittsburgh Cancer Institute Immunologic Monitoring and Cellular Products Laboratory (IMCPL) by trained personnel, according to standard operating procedures. The IMCPL is inspected by the College of American Pathologists (CAP) and the State of Pennsylvania, and certified by Clinical Laboratory Improvement Amendments (CLIA).

2.2. Luminex assay

Luminex multiplex assays of patient sera diluted 1/2 were performed in duplicate in 96-well filter-bottom microplates (Millipore, Billerica, MA) on 3/19/2009 or 3/24/2009. The microplate was blocked for 10 min with phosphate buffered saline/bovine serum albumin.

To generate a standard curve, 8-fold serial dilutions of appropriate standards were prepared in serum diluent. Luminex assays were performed according to manufacturer's instructions (Invitrogen, Camarillo, CA). Samples were analyzed using the Bio-Plex suspension array system (BioRad Laboratories). Analysis of experimental data for each analyte was done by fitting a standard curve with 8 parameters, as previously described (Gorelik et al., 2005). We used multiplex QC controls (with low, medium, and high analyte levels) from R&D Systems (Minneapolis, MN), a normal control plasma, and an IL-12p70 positive control prepared by the IMCPL. For the most commonly tested cytokines, the laboratory has established normal ranges.

The analytes tested and their lower limits of detection are as follows: IL-1 β (5 pg/ml), IL-1RA (15 pg/ml), IL-2 (3 pg/ml), IL-2R (32 pg/ml), IL-4 (3 pg/ml), IL-5 (11 pg/ml), IL-6 (3 pg/ml), IL-7 (5 pg/ml), IL-8 (2 pg/ml), IL-10 (7 pg/ml), IL-12p40/p70 (5 pg/ml), IL-13 (7 pg/ml), IL-15 (7 pg/ml), IL-17 (11 pg/ml), IFN- α (4 pg/ml), IFN- γ (6 pg/ml), TNF- α (3 pg/ml), G-CSF (22 pg/ml), GM-CSF (9 pg/ml), Eotaxin (2 pg/ml), IP-10 (1 pg/ml), MCP-1 (17 pg/ml), MIP-1 α (3 pg/ml), MIP-1 β (6 pg/ml), MIG (4 pg/ml), RANTES (4 pg/ml), EGF (6 pg/ml), FGFbasic (3 pg/ml), HGF (7 pg/ml), VEGF (4 pg/ml).

2.3. Statistical analysis

When analyte levels were below the lower limit of detection, they were set to zero; when above the upper limit of quantitation, they were set to that upper limit. Because data were highly skewed and contained zeros, they were categorized into quintiles; if there were too many zeros to allow quintiles, a smaller number of categories was used. Cox regression was used to examine the association of each analyte level with survival; inference was via the likelihood ratio test. Covariates associated with survival at level 0.1 were included in the Cox models. Spearman's test was used to assess correlation between pairs of variables.

3. Results

3.1. Baseline characteristics and summary statistics

Patients were 33 males and 23 females with a median age of 53 (range: 21 to 83), and Eastern Cooperative Oncology Group (ECOG) performance status as follows: 27 patients with ECOG=0; 27, ECOG=1; 2, ECOG=2. Forty patients received treatment for their disease; 16 did not. Patients' treatments varied, and included IFN- α IL-2, Temodar, and several other agents. Thirteen patients were classified as M1a melanoma (distant skin, subcutaneous tissue or lymph node involvement), 13 as M1b melanoma (lung involvement) and 30 as M1c melanoma (other visceral involvement). Thirty-eight patients died and 18 were censored; median survival was 9 months, and median censoring time was 28 months. Median storage time at -80 °C was 50 months, and ranged from 13 to 94 months.

3.2. Association of covariates with survival

We used univariate Cox models to assess the association of survival with age, sex, site of metastases (a categorical variable with 3 levels, noted above), ECOG performance status, and whether a patient was treated. (Due to the large variety of treatments and the small number of patients, we made no attempt to investigate effects of individual treatments.) Only site of metastases ($p=0.036$) and whether a patient was treated ($p=0.10$) had p -values sufficiently small that we felt they should be included in further analyses. The other covariates had p -values >0.23 .

3.3. Association of analytes with survival

The Luminex assays were done with the same lot number 30-Plex kit on 3/19/2009 and 3/24/2009. The association of each analyte with survival was assessed with Cox models that

included variables for site of metastases and whether a patient was treated. We found the levels of the following 17 analytes to be significantly associated with survival ($p < 0.05$): IFN- α ($p=0.00005$), IL-4 ($p=0.0004$), TNF α ($p=0.0006$), eotaxin ($p=0.0009$), G-CSF ($p=0.001$), IL-15 ($p=0.001$), MIP-1 α ($p=0.002$), IL-17 ($p=0.004$), IL-12p40p70 ($p=0.005$), IL-2 ($p=0.005$), IL-1Ra ($p=0.005$), IL-7 ($p=0.01$), MIP-1 ($p=0.01$), HGF ($p=0.01$), FGFbasic ($p=0.012$), IL-2R ($p=0.02$), and IL-13 ($p=0.03$). An analysis holding the false discovery rate to 10% picked up the same 17 analytes.

3.4. Red flags

This study was rather small (56 patients), and inter-patient variability of analyte levels was high. Thus, we felt that finding 17 of 30 analytes to be significantly associated with survival, some with exceptionally small p -values, was difficult to accept as a correct result. Thus, we viewed this finding as the first red flag. We also found that higher analyte concentrations were associated with prolonged survival for each of the 17 analytes; that was also hard to accept, and was thus the second red flag. (The 30 analytes included some (such as IFN- α) for which we hypothesized that elevated levels might be associated with longer survival, as well as others (such as IL-4) that we thought would not have that association.) Because of these red flags, we attempted to determine whether the results might be artifacts. We first looked at all possible (136) correlations between pairs of the 17 analytes using Spearman's rho and test. All correlations were positive (median correlation: 0.60), all but two were significant at $p < 0.05$, and most were very highly significant (median p -value: 6.5×10^{-6}). That finding (which was consistent with the second red flag) was the third red flag, and convinced us that the significant associations of the 17 analytes with survival must be some sort of artifact.

3.5. The underlying problems

We next looked at the relationship between storage time at -80°C and survival. It's reasonable to expect that sera from patients who survived longer would have been stored longer. Fig. 1 shows the relationship between storage time and survival time with the line of equality superimposed. Spearman tests of correlation of the two variables are significant ($p=0.0008$) for the censored patients (those alive at the time of the analysis), nearly significant for patients who died ($p=0.0504$). (We would expect that events for censored patients to be close to the line of equality if the date of the last follow up were close to the date of the assay.) The test is not significant for both groups combined ($p=0.53$); however, because the groups obviously differ, the test should not be used without stratifying on group. It is obvious from the figure that, for a given event time, sera from censored patients were almost always stored for less time than sera for patients who died; that can be expected to cause problems, as will be discussed further below.

Fig. 2 is a plot of IFN- α level versus storage time with a smooth curve superimposed. (Recall that IFN- α was the analyte most significantly associated with survival.) Correlation between the two variables is high (-0.67) and highly significant ($p=2.6 \times 10^{-7}$). The longer the time the sample was stored, the lower the IFN- α level, and the effect seems similar for censored and uncensored patients. Naively, one would think that would imply that lower levels of analyte would correspond to longer survival because longer survival is associated with longer storage time. But, this is the opposite of what was observed.

To resolve the apparent contradiction between the trend evident in Fig. 2 and the results of the analysis (that higher analyte levels were associated with longer survival), it is helpful to think in terms of the methodology used in the Cox model. Inference about the significance of the analytes is based on comparing the analyte level for each patient who died with the analyte levels of all patients who were still alive at the time of the patient death (or died at

the same time). Consider the death at the intersection of the vertical and horizontal lines in Fig. 1. The analyte level of that patient is compared with the levels of all patients (both censored and uncensored) on or to the right of the vertical line. Patients below the horizontal line have shorter storage times than that for the patient who died, and therefore will generally have higher analyte levels. All but 4 of the censored patients are below the horizontal line, but only 3 of the uncensored patients are. Therefore, the censored patients make it appear that higher analyte levels are associated with longer survival. Fig. 3 shows that censored patients have systematically higher analyte levels than do uncensored patients. It is the systematic difference in storage times between the censored and uncensored patients coupled with the apparent decrease in analyte level over time that led to the spurious association between survival and analyte level that we observed in the data.

If the censored patients are removed from the data, there is no association of IFN- α level with survival ($p=0.34$). Removing the censored patients biases the survival estimate, but will neither induce nor eliminate the association of a marker level with survival. Removing the censored patients seems justified in this instance because the characteristics of the censored patients systematically differ from those of the uncensored patients, but this procedure does adversely affect the power of any statistical tests. In our case, only about 1/3 of the patients were censored, so eliminating them from the analysis would be expected to reduce the significance of a very strong association ($p=0.00005$) but not eliminate it entirely if it were real.

4. Discussion

The most obvious explanation of trend apparent in Fig. 2 is that analyte levels decline over time even while stored at -80°C . We view this as a reasonable explanation because declines have been reported for some of the markers we measured (de Jager et al., 2009) but cannot rule out other explanations. For example, blood samples were obtained over a period of years during which sample handling protocols did not change, but different individuals responsible for the sample handling and processing procedures may not have followed the protocols in exactly the same manner. It is also possible that patient characteristics changed over time in some ways that we failed to identify (although it is hard to imagine that such changes could conspire to result in the sorts of changes that we see in Fig. 2 for 17 analytes). Finally, just as we were skeptical about the results of the original analyses, we are unwilling to draw conclusions about implications of Fig. 2 in the absence of further studies. We recently undertook studies of the stability of analyte levels over time (Butterfield et al., 2011), but we found that interpretation of the results was confounded by lot-to-lot variation in assay kits. Stability of analyte levels over time is a difficult and important problem, and additional studies are needed to confirm and characterize lack of stability. To protect against the effects of the possible decline of analyte levels over time in future biomarker studies, we recommend that storage time be as short and uniform as feasible until definitive studies of biomarker stability are completed.

One might question whether the pattern of censoring and deaths apparent in Fig. 1 is unique to our study. It is not. Consider the following elements of a common, but idealized design: 1) subjects are accrued over a period of time; 2) blood samples are collected and stored, and the survival clock is started at the time of accrual; and 3) subjects are followed until the time of the assays, which are done at the same time for all subjects. (Our study deviated somewhat from this, but for most subjects this was close to the design used.) In this situation, all censored subjects would be along the diagonal line (of equality) in Fig. 1, and all deaths would be above the line, similar to what we observe. In studies which use this basic design, the fraction of subjects that is censored will affect the results if analyte levels decline over time. With no censoring, increased levels would appear to be associated with

decreased survival. With sufficient censoring, increased levels would be associated with increased survival, and the greater the fraction of censored subjects, the stronger the apparent association of increased level with survival.

We could have attempted to model the analyte level decline over time and include that in the Cox regression, rather than remove the censored patients. However, typically censoring time is essentially equal to storage time, and deaths that occur for patients with samples stored for a short time are less than or equal to the storage time, so storage time and survival time must be correlated. Therefore, storage time is a very serious confounding variable, and we were concerned that an incorrect model could lead to false positive results. In addition, we were convinced that the elimination of any significant association of analyte level with survival upon removal of the censored patients was good evidence that there was no true association.

We noted that blood was drawn after diagnosis, but before treatment. Typically, blood was drawn within weeks after diagnosis, but that was not always the case: in some cases, blood was drawn several months after diagnosis. Had this study yielded statistically significant results, patients with a large delay between diagnosis and blood draw would have been eliminated because their analyte levels might be a poor proxy for the baseline levels, and because they could introduce bias. The bias would arise if analyte levels changed between diagnosis and blood draw, and this would occur if levels changed with disease progression. For example, suppose one group (A) of patients had blood drawn 6 months after diagnosis, and another group (B) had blood drawn within weeks after diagnosis. The two groups would have different analyte levels. Patients in group B could die within 6 months, but those in group A could not. Therefore, group A's analyte levels would be associated with longer survival than group B's.

5. Conclusions

Retrospective studies of all sorts have notorious reputations because these studies can have undetected biases that result in invalid conclusions. Even carefully designed epidemiological studies have yielded results that could not be reproduced by randomized clinical trials; a well-known example is the relationship of the intake of β -carotene to lung cancer (Omenn et al., 1994). Biomarker studies have an even worse track record: "Despite years of research and hundreds of reports on tumor markers in oncology, the number of markers that have emerged as clinically useful is pitifully small. Often, initially reported studies of a marker show great promise, but subsequent studies... yield inconsistent conclusions or stand in direct contradiction...." (Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics, 2005). Therefore, it is critical that investigators view their results with skepticism before claiming new findings. That skepticism should lead to detailed examination of the data with the goal of ruling out (or finding) alternative explanations for the results.

Acknowledgments

We would like to thank Sharon Sember and UPCI Immunologic Monitoring Laboratory for performing the LuminexTM multiplex assays in this study, and are grateful for the comments of two anonymous reviewers. This study was supported in part by the University of Pittsburgh Cancer Institute and the NIH Cancer Center Support Grant P30 CA047904; NCI RO1 CA138635 (LHB); Developmental Research Funds of the SPORC in Skin Cancer P50 CA121973 (JMK); Frontier Science and Technology Research Foundation and ECOG Central Laboratory Support (LHB).

References

American Cancer Society, y. Cancer facts and figures. 2006. <http://www.cancer.org/downloads/STT/CAFF2006PWSecured.pdf>

- Balch, CM. Cancer: Principles and Practice of Oncology. 5. Lippencott; Philadelphia: 1997.
- Balch CM, Soong SJ, Gershenwald JE, et al. Prognostic factors analysis of 17,600 melanoma patients: validation of the American Joint Committee on Cancer melanoma staging system. *J Clin Oncol*. 2001; 19:3622. [PubMed: 11504744]
- Butterfield LH, Potter DM, Kirkwood JM. Technical and biostatistical issues in multiplex serum biomarker assessments. *J Transl Med*. 2011; 9:173. [PubMed: 21989127]
- de Jager W, Bourcier K, Rijkers GT, et al. Prerequisites for cytokine measurements in clinical trials with multiplex immunoassays. *BMC Immunol*. 2009; 10:52. [PubMed: 19785746]
- Gorelik E, Landsittel DP, Marrangoni AM, et al. Multiplexed immunobead-based cytokine profiling for early detection of ovarian cancer. *Cancer Epidemiol Biomarkers Prev*. 2005; 14:981. [PubMed: 15824174]
- Kim CJ, Reintgen DS, Balch CM. The new melanoma staging system. *Cancer Control*. 2002; 9:9. [PubMed: 11907461]
- Korn EL, Liu PY, Lee SJ, et al. Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials. *J Clin Oncol*. 2008; 26:527. [PubMed: 18235113]
- Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics: reporting recommendations for tumor marker prognostic studies. *JNCI*. 2005; 97:1180. [PubMed: 16106022]
- Omenn GS, Goodman G, Thornquist M, et al. The β -Carotene and Retinol Efficacy Trial (CARET) for chemoprevention of lung cancer in high risk populations: smokers and asbestos-exposed workers. *Can Res Suppl*. 1994; 54:2048s.
- Thompson JF, Scolyer RA, Kefford RF. Cutaneous melanoma. *Lancet*. 2005; 365:687. [PubMed: 15721476]

Fig. 1. Scatter plot of analyte storage time at -80°C vs. event time (time of death or time of censoring). The point for a patient with equal storage and event times would fall on the diagonal line. See text for description of the horizontal and vertical lines.

Fig. 2.
Scatter plot of IFN- α level versus storage time with a smooth curve superimposed.

Fig. 3.
Scatter plot of IFN- α level versus event time.