RESEARCH ARTICLE

Clinical and laboratory factors contributing to uninterpretable beryllium lymphocyte proliferation tests (BeLPT)

Derek E. Smith MS[1] (iD) | Ashley P. Golden PhD[2] (iD) | Arthur W. Stange PhD[3] |
Elizabeth Barker MPH[4] | Margaret Mroz MSPH[4] | Anna E. Barón PhD[5] |
Debashis Ghosh PhD[5] | Lisa Maier MD[4] | Donna Cragle PhD[2] |
Lee S. Newman MD, MA[6]

[1] Cancer Center Biostatistics Core, University of Colorado, Anschutz Medical Campus, Aurora, Colorado

[2] Oak Ridge Associated Universities, Oak Ridge, Tennessee

[3] Oak Ridge Associated Universities, Arvada, Colorado

[4] National Jewish Health, Denver, Colorado

[5] Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado, Anschutz Medical Campus, Aurora, Colorado

[6] Center for Health, Work, and Environment, and Department of Environmental and Occupational Health and Department of Epidemiology, Colorado School of Public Health, University of Colorado, Anschutz Medical Campus, Aurora, Colorado

Correspondence
Derek E. Smith, MS, AMC Building 500, RM N2215, Mail Stop C243 13001 E 17Th Place Aurora, 80045 CO.
Email: derek.e.smith@ucdenver.edu

Funding information
Oak Ridge Associated Universities

Background: The beryllium lymphocyte proliferation test (BeLPT), has become the principal clinical test for detecting beryllium sensitization and chronic beryllium disease. Uninterpretable BeLPT results can occur in a small but significant proportion of tests from poor lymphocyte growth (PG) or over proliferation of lymphocytes (OP). The clinical and laboratory causes of uninterpretable results are not known.

Methods: BeLPT data from the US Department of Energy-supported Former Worker Screening Program were analyzed for a 10-year period. Drivers of uninterpretable BeLPTs were investigated using multivariable models and classification techniques.

Results: Three participant attributes were significantly associated with PG, while OP showed no significant associations. Serum lot for the lymphocyte growth medium accounted for 21% of the variation in PG and 16% in OP.

Conclusion: Serum lots influence the likelihood of having uninterpretable BeLPT. To better understand uninterpretable results and possibly reduce their occurrence, additional laboratory-related factors should be addressed.

KEYWORDS
BeLPT, BeS, CBD, chronic beryllium disease, lymphocyte proliferation

1 | INTRODUCTION

Exposure to beryllium particulate, aerosols, or fumes, usually in occupational settings, may result in beryllium sensitization (BeS), an adaptive immune response. BeS can progress to chronic beryllium disease (CBD), an incurable, debilitating lung condition. Individuals with BeS develop CBD at a rate of 3.2-9.2% per year and prevalence rates of BeS and CBD for workers exposed to beryllium range widely depending on the industry.[1] Prevalence rates between 0.9 and 11.8% for BeS and between 0.09 and 7.8% for CBD have been noted, with nuclear weapons and beryllia ceramics facilities accounting for some of the highest rates.[2]

The blood beryllium lymphocyte proliferation test (BeLPT) was developed as an in vitro clinical and medical screening test to identify BeS, which is one step in diagnosing CBD.[3] With 68.3% test sensitivity and 96.9% test specificity in medical surveillance programs, the BeLPT

has proven to be the most reliable, non-invasive method for identifying BeS.[4] There are a number of intricacies involved with the interpretation of the BeLPT test results.[5–8] Ideally, BeLPT results are reported as normal, abnormal, or borderline depending on the proliferative response to beryllium. Yet, approximately 4% of participants undergoing the BeLPT can expect to receive a test result of "uninterpretable."[4] Uninterpretable results for this cell culture-based assay have been hypothesized to be due to performance of the blood test in the laboratory or participant specific risk factors such as their underlying health conditions.

Results of the BeLPT, like other lymphocyte proliferation assays, are reported as a "stimulation index," which is a simple ratio of the amount of lymphocyte proliferation seen in cells that have been cultured in the presence of a mitogen or antigen (such as beryllium) divided by the amount of lymphocyte proliferation seen in cells that were placed in the same culture conditions, but with no stimulus. From a laboratory perspective, potential causes of uninterpretable results may be traced to problems with control wells or to the response of cells that have been stimulated with beryllium, antigens, and mitogens. Cells in unstimulated control wells may demonstrate poor lymphocyte growth (which risks falsely exaggerating the stimulation index by depressing the denominator), or may spontaneously over-proliferate (which risks falsely underestimating the stimulation index, by inflating the denominator). Uninterpretable results can also occur when cells show a poor response to positive controls (ubiquitous mitogens and antigens to which all lymphocytes should respond), increased cell death due to problems with blood transport or cell culture conditions, or due to other technical and equipment-related problems that can produce unacceptably high coefficients of variation. As a result of these factors, though the occurrence is infrequent, further refinement of the conditions affecting the BeLPT is warranted.

Although unproven, it is plausible that individual variation in lymphocyte function may produce uninterpretable BeLPT results. Underlying health conditions, such as hematopoietic malignancies, autoimmune disorders, and immunomodulatory medications, such as corticosteroids, may cause lymphocytes to either over-respond or under-respond in vitro.

We sought to investigate two of the primary reasons that a BeLPT is uninterpretable, which included poor lymphocyte growth (PG) and over-proliferation (OP) of lymphocytes in the control wells during the BeLPT. Specifically, we sought to investigate BeLPT classification of uninterpretable results and their association with participant specific factors. Cell growth is known to be affected by the specific serum lot used in the culture medium, with some lots supporting more or less robust proliferation in response to antigens and mitogens.[8] In addition, procedural differences in the BeLPT can occur across laboratories and are important factors to control for when the goal is to assess whether there are any participant specific risk factors associated with an uninterpretable test result.

Under newly accepted rules by the Occupational Safety and Health Administration (OSHA), beryllium screening will increase and the BeLPT will be used as the standard screening method. Current workers meeting the criteria set forth in OSHA's occupational exposure to beryllium rule will be required to have a BeLPT as part of medical surveillance and follow up BeLPTs every 2 years.[9] Insight into PG and OP test results may help clinicians explain test results to workers and identify whether uninterpretable results are masking a true relationship between beryllium exposure and an individual's immune response. Furthermore, lymphocyte proliferation tests are used in a number of clinical settings (eg, to assess for immunocompetence and hypersensitivity). For example, lymphocyte proliferation assays have been investigated for their potential to diagnose individuals with hypersensitivity to solvents, disinfectants, and other metals.[10–14] While the data used in this study focus on use of lymphocyte proliferation in surveillance for beryllium health effects, the findings may prove applicable to other lymphocyte activation/proliferation tests that can also occasionally yield uninterpretable results.

2 | METHODS

2.1 | Study population

Demographic and underlying health condition information was obtained through medical examination data collected by the National Supplemental Screening Program (NSSP). The NSSP is a nationwide occupational medical screening program that is part of the larger U.S. Department of Energy's (DOE) Former Worker Medical Screening Program (FWP). The FWP, a DOE-funded program initiated in 1996, was designed to identify adverse health outcomes related to hazardous occupational exposures for former DOE site workers.[15]

The FWP offers free medical screening examinations to all former DOE workers, contractors, and subcontractors. The NSSP began medical examinations for former DOE site workers in 2005. At the time of this publication, the NSSP had collected medical data for over 15 000 former DOE workers. If participants in the NSSP indicate they may have been exposed to beryllium in the workplace, or if information on site exposures indicates that there was potential beryllium exposure, the BeLPT is administered as part of the medical screening examination. Accordingly, the population investigated in this study represents former DOE workers who were enrolled in the NSSP and have had an occupational exposure to beryllium due to their work with or around beryllium metal, alloys, and or ceramics. Workers at least 17 DOE facilities currently have or have had the potential for beryllium exposure, and between 54 000 to 134 000 current workers have potential exposure to beryllium in the workplace.[16,17]

As part of the NSSP, former DOE site workers are eligible to voluntarily complete an initial medical examination. Every 3 years, individuals become eligible for a rescreening examination. Components of the medical examination include a physical examination, self-reported health history, DOE work exposure survey, basic blood tests, a chest X-ray, hearing tests, pulmonary function, fecal occult blood, and a urinalysis. Depending on an individual's work history, further exposure-specific tests may be conducted, for example, the blood BeLPT. Once individual informed consents are obtained, blood samples are analyzed at one of two BeLPT

laboratories: National Jewish Health (Denver, CO) or Oak Ridge Institute for Science and Education Beryllium Testing Laboratory (Oak Ridge, TN).

This study assessed NSSP participants whose BeLPTs were performed by either laboratory between October 2005 and December 2015 ($n = 18\,321$ total tests). The investigators neither recruited participants nor collected new data for this study. All data were de-identified by NSSP staff before being sent to the investigators. Thus, IRB approval was obtained with exempt status, with additional informed consent not required.

2.2 | Definition of PG and OP

In vitro lymphocyte growth requires the addition of pooled human AB-positive serum to the growth medium (RPMI 1640) during the incubation period of the BeLPT.[3] Since cell growth is affected by the specific serum lot the threshold at which cultures experience OP is specific to a given pooled serum lot and has been routinely defined as follows. The OP serum specific cut-point is created by performing BeLPTs on roughly 30 blood samples from individuals who have no prior history of beryllium exposure. The OP cut-points are defined as the 99.9 percentile of the cell growth in control samples, that is, cells that have been incubated in growth media in the absence of beryllium salts. Alternatively, the PG cut-points are defined as twice the background noise (ie, counts with growth media alone). However, these definitions are not shared across laboratories and the information needed to apply these methods to the entire BeLPT dataset was not available. Thus, for purposes of this study, the OP and PG cut-points were redefined as follows in order to harmonize the data across the laboratories.

Eleven different pooled serum lots were used to run the BeLPTs for the NSSP over the 10-year study period. BeLPT interpretation occurs at two different time points for cells in culture. However, these interpretation times were not the same in both labs. One lab interprets the BeLPTs after 4 and 6 days of incubation and the other after 5 and 7 days. Thus, OP thresholds were defined for two separate interpretation times for a given serum lot. In order to determine the OP threshold, we identified individuals who only ever had a normal BeLPT test result, referred to from here on as "exposed normals." For individuals with multiple BeLPTs, their first test result was chosen. Cell counts from the control wells were obtained for this group of exposed normal individuals. These test results were then stratified by serum lot, and for a given serum and interpretation time (eg, Day 4), an OP threshold was defined as follows. First, for a given serum lot and interpretation time, 25% of the population of exposed normal individuals was randomly sampled. A sensitivity analysis was used to identify the percentage of exposed normal individuals to be randomly sampled that yielded a representative sample, yet did not oversample the population. An OP threshold was defined for a given serum and interpretation time by obtaining the 99.9 percentile of the cell counts from the random sample of exposed normal individuals. This process was repeated for each serum and interpretation time. Unlike OP, a fixed value of less than or equal to 85 counts per minute

(CPM) of cells in control wells was defined as the PG cut-point. This value was determined by clinical and laboratory subject matter experts.

2.3 | Case-control selection

BeLPT test results were excluded from the case-control selection if the BeLPT result was abnormal ($n = 617$), borderline ($n = 416$), or invalid due to reasons not related to lymphocyte growth ($n = 726$). The main reasons for invalid tests were attributed to cell death prior to culture or the quantity of the blood sample not being sufficient to run the BeLPT. The remaining BeLPTs were classified as cases (ie, PG or OP) using the following routine definition. A test result was classified as OP if cell growth for an individual's control wells was greater than the threshold at both interpretation times, or if cell growth was more than 10% higher than the threshold at one of the interpretation times. Similarly, a test result was classified as PG if cell growth for an individual's control wells was less than the threshold at both measurement times, or if cell growth was more than 10% lower than the threshold at one of the interpretation times. There were a few instances (<5) in which individuals with multiple BeLPTs in the NSSP dataset were classified as both PG and OP. These individuals were removed from the study. There were 1831 individuals who had underwent multiple NSSP screenings. Those with multiple test results could have had multiple uninterpretable of the same type (ie, PG or OP), a mixture of normal and uninterpretable results, or all normal results. In order to be consistent with how individuals with multiple BeLPTs were treated, in the event one had multiple BeLPTs and one or more was PG or OP, the initial uninterpretable test result was used. The same was true for those with multiple normal BeLPTs, in which case results from their initial BeLPT were used. Lastly, we defined controls (ie, individuals with normal CPM in their cell cultures) as the remaining individuals who were not classified as OP or PG, and who had never had an abnormal or borderline test result. Generally, an individual's BeLPT date and clinical screening date coincided, however this was not always true. Thus, clinical screening data were only used if the data were captured within a 6-month period surrounding the BeLPT date.

2.4 | Measures

All participants' clinical variables obtained from the NSSP questionnaire and medical examination data that had enough data ($n = 174$) were investigated as potential predictors of the BeLPT outcomes. The NSSP does not collect information on all variables for each participant. Thus, variables that had less than 20 responses for a given outcome (ie, PG or OP) were excluded from the analysis due to a lack of data to make a meaningful assessment. In order to assess the association of the different NSSP variables with the outcome of interest (ie, PG or OP), two different approaches were taken. First, we relied on subject matter experts to identify biologically important variables. Yet, due to the gap in our understanding about what risk factors may be associated with PG and OP, we also used a univariate approach to identify potential candidate variables that were highly associated with a given

outcome. This was an attempt to reduce the risk that we overlooked important variables. The two approaches are described in detail below.

2.5 | Statistical analyses—Association

Both the subject matter expert approach and the univariate association approach employed a multivariable generalized linear mixed model with a binary link function. This enabled us to model the relationship between the outcome of interest and the selected covariates, while accounting for the correlation between tests that were performed using the same serum lot. The proportion of the total variance in the outcome that can be attributed to serum lot was assessed using the intraclass correlation coefficient (ICC). In addition, since the NSSP dataset is primarily composed of older males, all models were adjusted for gender and age. Age was treated as a categorical variable and separated into four categories; <55, 55-64, 65-74, and 75 years and over. Lastly, it was necessary to control for the lab where the test was performed in all models by including it as a covariate. This was done on the basis that test performance can vary across labs and was necessary to account for when assessing the association between uninterpretable BeLPTs and participant characteristics.

Mixed models were run using PROC NLMIXED in SAS version 9.4 and glmer from the lme4 package in R version 3.2.4.[18–20] Linearity on the logit scale between individual variables and the outcome of interest was assessed for continuous variables. Continuous variables found to display a non-linear relationship with the outcome were categorized based on laboratory interpretations to normal/abnormal or low, normal, or high.

The first approach selected candidate variables for the multivariable model that were chosen by subject matter experts (five in total). Subject matter experts were asked to rank the candidate variables on a scale of 0-5 (five representing the strongest relationship) for how related they expect them to be with each outcome. A given variable was then ranked based on its combined score (maximum score of 25). Candidate variables were identified as variables with a combined score greater than 10 and included in the multivariable model. Collinearity was assessed by the variance inflation factor (VIF) and any variables with a VIF greater than 5 were investigated further. In the event variables were collinear, the one with the lower ranking was dropped. Interrater reliability was assessed using Krippendorff's Alpha, where an alpha value of 1 indicates perfect agreement and a value of 0 indicates agreement that is no better than chance.[21]

The second approach selected candidate variables by assessing their association with a given outcome in the univariate setting. Variables found to have a P-value less than 0.05 were considered as candidate variables for the multivariable model. Collinearity issues were assessed using the criteria previously discussed. However, in the event variables were collinear, variables with a weaker association in the univariate setting were removed in favor of keeping those with a stronger association.

In order to achieve a parsimonious model, variable reduction for the multivariable model was conducted using an all-subsets approach, implemented using the glmulti package in R.[22] Essentially, for each approach described above all possible model subsets of the candidate variables were modeled and ranked using the corrected Akaike's information criterion (AICc). The model with the lowest AICc out of the two approaches was chosen and interpreted for a given outcome. Odds ratios (OR) from the resulting model were used to assess the association between the outcome and individual predictors. In order to account for multiple testing in the multivariable models, the false discovery rate (FDR) and corresponding adjusted P-values were computed (FDR maintained at 5%) based on the number of covariates assessed in the model.[23] Accordingly, significance of the model results was assessed based on the results of the FDR adjusted P-values.

2.6 | Statistical analyses—Classification

The NSSP dataset provided us with the unique opportunity to explore alternative approaches when attempting to use participant specific factors to predict an uninterpretable BeLPT result. Due to the size of the NSSP dataset, in addition to using a multivariable generalized linear mixed model approach, we also explored the use of the machine learning technique called classification and regression tree (CART). This also provided us with a means to validate whether the conclusions were similar among the different methods. The dataset was split into a training set (2/3) and test set (1/3) and missing data were imputed using a single K Nearest Neighbors (KNN) approach ($K = 10$) implemented using the DMwR R package.[24,25] It is often vital to estimate the uncertainty in imputed data using multiple imputation when the objective is estimation, yet this is not a necessity when the objective is prediction.[26] Variables with a large amount of missingness (ie, over 1000 missing records) were not imputed and were instead excluded from the classification analysis.

Whether an individual could be classified as PG and OP was evaluated using both classification and regression tree (CART) and multivariable generalized linear mixed model approaches. These two approaches were explored since the multivariable generalized linear mixed model framework is limited by multicollinearity, the number of predictors that can be included without overfitting the model, and the inclusion of variables with little to no variation (near 0-variance). The CART framework is much less susceptible to these issues.[27]

A CART can be constructed using a number of techniques. We chose to explore two of these, recursive partitioning for classification (RPART) and the cost C5.0 algorithm where boosting was used to combine the ensemble of classification trees to make predictions. Boosting was implemented using the AdaBoost algorithm for RPART and C5.0 boosting for the C5.0 algorithm.[27] The boosted classification trees were run using the caret package in R version 3.2.4.[28] These algorithms are discussed in detail in the book Applied Predictive Modeling.[27]

Models were ranked/evaluated using partial area under the receiver operating characteristic curve (AUC), where specificities were only assessed in the interval where sensitivities ranged from 0.5 to 1.0. The partial AUC was used over the complete AUC since we were interested in evaluating different models that had high sensitivity. In the case of screening for PG and OP, an individual will receive the

BeLPT regardless of their predicted result. Thus, a false positive (ie, incorrectly predicting one as having a PG or OP test result when in fact their BeLPT came back normal) does not result in any unnecessary testing. Yet, a false negative (ie, incorrectly predicting one as having a normal test result when in fact their BeLPT came back either PG or OP) provides no insight into why the BeLPT was uninterpretable. Therefore, allowing for a higher false positive rate enables us to capture more true positives while having no adverse consequences on patients. Finally, the partial AUC values were rescaled to yield the typical interpretation of complete AUC values between 0 and 1 using the pROC package in R version 1.10.0.[29,30]

In order to account for correlation across BeLPTs that were run using the same serum lot, a random intercept was included in the multivariable mixed model. Yet, when using CART to assess OP and PG it was not possible to account for correlation within serum lot in this manner. Instead, we included empirical Bayes estimates of the random intercepts for serum lot as a continuous predictor in the CART analyses. This was accomplished using PROC NLMIXED in SAS version 9.4 by running a univariate generalized linear model with a binary link function on the outcomes of interest (ie, PG and OP) where only serum lot was included in the model as a random intercept. Empirical Bayes estimates where then obtained for the random intercepts and transformed to OR.

The CART models were trained over a number of tuning parameters using the training dataset and bootstrapping with 50 replicates. AdaBoost models were tuned using two different versions of the algorithm (M1 and Real) and up to 150 boosting iterations.[31] The boosted C5.0 models were tuned using up to 100 boosting iterations, whether the splits were tree or rule-based, and whether or not winnowing (ie, removal of unimportant predictors) should be applied.[26] For a given algorithm (AdaBoost or C5.0), the training model with the highest partial AUC was selected as the final model and its prediction was evaluated using the test dataset. For each final model the mean partial AUC and 95% confidence interval (z-based) was estimated using the test dataset and bootstrapping with 1000 replicates. Variable importance for the final CART models was assessed by determining the percentage of all tree splits that was associated with a given variable in the test dataset.

The multivariable generalized linear mixed model employed the same all-subsets approach that was used to assess the association of participant characteristics with PG and OP. The differences were that the partial AUC was used to rank models over the AICc and that the all-subsets approach was performed on the imputed dataset so that the datasets used in the classification analyses were identical.

Lastly, logistic regression may underestimate the probability of rare events.[32] Since our outcomes were just under 2%, we assessed whether predictions improved as the frequency of the outcomes increased. This was accomplished using Synthetic Minority Over-sampling Technique (SMOTE), which uses a combination of under-sampling the majority class (normal test results) and over-sampling the minority class (PG or OP) approaches to simulate new datasets that have a greater frequency of events from the minority class.[33] Using SMOTE, implemented using the DMwR package in R, the number of

cases were doubled (over-sampling) with various combinations of under-sampling to create three simulated datasets where the frequency of the cases was increased from roughly 2% to 10%, 20%, and 30%.[25] Simulated datasets were assessed using the same classification approaches detailed above that were used on the full unsimulated dataset.

3 | RESULTS

As of December 2015, 13 338 individuals had been screened with the BeLPT at least once through the NSSP. A total of 18 321 BeLPT records existed, which included repeat tests. The BeLPT records were composed of 7 507 tests that were completed at Oak Ridge Institute for Science and Education and 10 814 at National Jewish Health. As detailed in the methods section, abnormal, borderline, and invalid (ie, insufficient sample collected or nonviable cells) BeLPT test results were excluded. This provided us with a final cohort of 12 316 unique individuals with BeLPTs (6 804 from National Jewish Health and 5 512 from Oak Ridge Institute for Science and Education). Of the 12 316 BeLPT tests, 11 937 were normal BeLPTs, 154 were PG, and 225 were OP.

The assessment of whether the clinical variables had potential clinical and or scientific associations with PG or OP from subject matter experts yielded a low interrater reliability, Krippendorff's Alpha was 0.23 for PG and 0.25 for OP. Due to the differences in the backgrounds among the subject matter experts, interrater reliability was also assessed separately between clinical and laboratory/test experts. While clinical experts had slightly higher agreement (Krippendorff 's Alpha 0.29 for PG and 0.33 for OP) than laboratory/test experts (Krippendorff 's Alpha 0.18 for PG and 0.20 for OP), the interrater reliability remained low regardless of the area of expertise.

3.1 | Association for PG

An overview of the final cohort used in the multivariable model to test the association between PG and participant specific factors can be found in Table 1. Particular serum lots and labs were found to contain a large number of the PG results, further emphasizing the need to control for these variables when assessing participant specific risk factors (Table I). The model using the variables selected from the univariate approach was found to have the lowest AICc (1114.55) compared to the model resulting from subject matter experts approach (AICc: 1166.21). Ten variables were included in the all-subsets selection process for the univariate approach since they were shown to have a high association with PG in the univariate setting. Six of the initial 10 variables were retained in the final model and the results from the multivariable model are shown in Table 2.

The variables included in the multivariable model displayed a high association with PG in the univariate setting (differences between the groups are shown in Table 1) yet many of these variables no longer displayed a strong association with PG when included together in the multivariable model. The odds of PG were significantly different

TABLE 1 Participant characteristics across the poor growth and normal growth groups

	Outcome			
	Normal		Poor growth	
	N = 10094		N = 115	
Mean corpuscular hemoglobin concentration N (%)				
Normal	8756	(86.7)	78	(67.8)
Low	1338	(13.3)	37	(32.2)
White blood cell count N (%)				
Low	286	(2.8)	4	(3.5)
Normal	9524	(94.4)	102	(88.7)
High	284	(2.8)	9	(7.8)
Red blood cell count N (%)				
Low	1036	(10.3)	21	(18.3)
Normal	8711	(86.3)	88	(76.5)
High	347	(3.4)	6	(5.2)
Creatinine N (%)				
Normal	8523	(84.4)	77	(67.0)
High	1571	(15.6)	38	(33.0)
Hemoglobin (mg/dL) [mean (sd)]	14.52	(1.46)	13.68	(1.69)
Joint pain N (%)				
No	4995	(49.5)	69	(60.0)
Yes	5099	(50.5)	46	(40.0)
Lab N (%)				
1	4374	(43.3)	94	(81.7)
2	5720	(56.7)	21	(18.3)
Gender N (%)				
Males	7688	(76.2)	88	(76.5)
Females	2406	(23.8)	27	(23.5)
Age N (%)				
<55	1968	(19.5)	11	(9.6)
55-64	2810	(27.8)	32	(27.8)
65-74	3094	(30.7)	28	(24.3)
75	2222	(22.0)	44	(38.3)
Serum N (%)				
1	509	(5.0)	5	(4.3)
2	935	(9.3)	0	(0.0)
3	921	(9.1)	1	(0.9)
4	1300	(12.9)	13	(11.3)
5	770	(7.6)	5	(4.3)
6	1115	(11.0)	6	(5.2)
7	1978	(19.6)	67	(58.3)
8	728	(7.2)	2	(1.7)
9	772	(7.6)	16	(13.9)
10	85	(0.8)	0	(0.0)
11	981	(9.7)	0	(0.0)

between those with low versus normal mean corpuscular hemoglobin concentration (MCHC) (FDR adjusted P-value: 0.048). On average, the odds of PG were around 50% lower for individuals with normal MCHC compared to individuals with low MCHC (OR: 0.509, 95%CI: 0.299-0.868). High levels of MCHC were not examined because of the small number of individuals in this category. Hemoglobin concentration was also found to be significantly associated with PG (FDR adjusted P-value: 0.040) and displayed a similar relationship as seen with MCHC. On average, for every unit increase in hemoglobin, the odds of PG were 28% percent less likely (OR: 0.743, 95%CI: 0.619-0.891). Lastly, we found a significant difference in the odds of PG between those with joint pain versus those without joint pain (FDR adjusted P-value: 0.048). On average, for those who experienced joint pain, the odds of PG was 43% less likely compared to those that did not experience joint pain (OR: 0.569; 95%CI: 0.367-0.883). Notably, in addition to the exanimation of the clinical factors related to PG, laboratory serum lot was found to account for a sizable portion of the variation in PG with an ICC of 0.21.

3.2 | Association for OP

An overview of the final cohort used in the multivariable model to test the association between OP and participant specific factors can be found in Table 3. Again, particular serum lots and labs were found to contain a large number of the OP results, further emphasizing the need to control for these variables when assessing participant specific risk factors (Table 3). The model using the variables selected by the univariate approach was found to have the lowest AICc (1610.80) compared to the model using variables selected by subject matter experts (AICc: 1622.2). Ten variables were included in the all-subsets selection process for the univariate approach since they were shown to have a strong association with OP in the univariate setting. Five of these variables were retained in the final model. The multivariable model results can be found in Table 4. Similar to what was seen in PG, the variables included in the multivariable model displayed a high association with OP in the univariate setting (differences between the groups are shown in Table 3), yet when included together in the multivariable model no significant associations were found. In addition, akin to PG, we observed that a considerable portion of the variation in OP was attributed to serum lot, with an ICC of 0.16.

3.3 | Classification

A total of 167 different variables from the NSSP medial screenings were assessed using the CART analysis. The entire cohort of individuals was utilized for the classification analyses since missing data values were imputed using KNN. Thus, the cohort used to assess PG consisted of 12 091 individuals, including 11 937 controls (ie, normal BeLPT) and 154 cases (ie, PG). Similarly, the cohort used to assess OP consisted of 12 162 individuals, including 11 937 controls (ie, normal BeLPT) and 225 cases (ie, OP). Using the (2/3) training set and (1/3) test set split, the same number of normal tests were included in the training (7958) and test (3979) sets for PG and OP analyses.

TABLE 2 Multivariable model results used to assess the association for Poor Growth results

Model	OR	95%CI	Test statistic	P-value	FDR adjusted P-value
Poor growth					
Mean corpuscular hemoglobin concentration (normal vs low)	0.509	0.299-0.868	t_{10} = −2.82	0.018	0.048*
White blood cell count			$F_{(2,10)}$ = 2.5	0.132	0.170
Low vs normal	1.248	0.390-3.993			
High vs normal	2.294	0.993-5.295			
Creatinine (high vs normal)	1.547	0.933-2.564	t_{10} = 1.92	0.083	0.166
Red blood cell count			$F_{(2,10)}$ = 2.32	0.149	0.170
Low vs normal	0.734	0.356-1.515			
High vs normal	2.533	0.899-7.135			
Joint pain (yes vs no)	0.569	0.366-0.883	t_{10} = −2.86	0.017	0.048*
Hemoglobin concentration (mg/dL)	0.743	0.619-0.891	t_{10} = −3.64	0.005	0.040*
Gender (female vs male)	0.717	0.407-1.264	t_{10} = −1.31	0.22	0.22
Age			$F_{(3,10)}$ = 2.68	0.104	0.166

OR, Odds Ratio; Confidence Interval, CI; FDR, False Discovery Rate.
Results are shown for the model with the lowest AICc.
Sample sizes for the model can be found in the participant characteristics overview for Poor Growth in Table 1.
*indicates significance after accounting for multiple testing using the FDR

There were 103 cases in the training set and 51 in the test set for PG, and 105 cases in the training set and 75 in the test set for OP.

3.4 | Classification of PG

The majority of the top models from the multivariable linear mixed model using the univariate and all-subsets approach failed to converge. In contrast, the subject matter experts and all-subsets approach did not experience the same convergence issues. However, there was no indication that differences in the variables chosen between the two methods made a notable difference in the ability to classify an individual as PG. This was inferred by the range of partial AUC values for the top models that did converge from the univariate approach, which were comparable to the results from the subject matter experts approach (Univariate partial AUC: 0.774-0.779, Subject Matter Experts partial AUC: 0.75-0.766). As such, the results from the subject matter experts for the multivariable generalized linear mixed model are presented here since the selected variables did not result in model convergence issues.

The AdaBoost algorithm produced similar results across the different tuning parameters, with the M1 algorithm performing better than the Real algorithm in all cases (Figure 1). The C5.0 algorithm stabilized after 25 boosting iterations (Figure 2). Once stabilized, both algorithms displayed a similar range of partial AUC values (0.60 to 0.73) and an increase in oversampling coincided with an increase in the partial AUC in the training dataset.

Due to the similar performance seen across the three methods used to predict an individual as PG, we only present an example of the resulting ROC curves for the multivariable mixed model (Figure 3). Table 5 displays the partial AUCs resulting from the final models applied to the test dataset. Variable importance was found to be low across all

levels of oversampling, the number of splits associated with a given variable for AdaBoost ranged from 0.0 to 1.9% and 0.0 to 8.0% for C5.0.

3.5 | Classification of OP

The results from the multivariable generalized linear mixed model using the variables selected from the univariate approach yielded the highest partial AUC. The range of partial AUC values from the top models of the all-subsets approach was slightly higher for the univariate approach. Yet, there was not a large difference in the partial AUC values between the two approaches (Univariate partial AUC: 0.676-0.681, Subject matter experts partial AUC: 0.656-0.668).

The overall trends in the training results for OP were similar to what we observed for PG. The AdaBoost M1 algorithm performed better than the Real algorithm in all cases, and the C5.0 algorithm was fairly stable after around 25 boosting iterations the (Figure 1 and 2). Both algorithms displayed a similar range of partial AUC values (0.53 to 0.60) after they stabilized, and an increase in oversampling coincided with a slight increase in the partial AUC.

The performance across the three methods (measured by the partial AUC) was similar but slightly worse than PG, examples of the resulting ROC curves are not shown here for brevity. The partial AUCs resulting from the final models applied to the test dataset are shown in Table 5. Again, variable importance was low, and across all levels of oversampling, the number of splits associated with a given variable for AdaBoost ranged from 0.0 to 1.8% and 0.0 to 6.0% for C5.0.

4 | DISCUSSION

Uninterpretable BeLPTs are an infrequent, but potentially costly, occurrence in large scale beryllium disease surveillance programs. If

TABLE 3 Participant Characteristics across the over proliferation and normal growth groups

	Outcome			
	Normal		Over proliferation	
	N = 9577		N = 165	
Lymphocyte count N (%)				
Low	346	(3.6)	2	(1.2)
Normal	9139	(95.4)	157	(95.2)
High	92	(1.0)	6	(3.6)
Burning pain in hands/feet N (%)				
No	8133	(84.9)	127	(77.0)
Yes	1444	(15.1)	38	(23.0)
Numbness/tingling in fingers/toes N (%)				
No	6072	(63.4)	87	(52.7)
Yes	3505	(36.6)	78	(47.3)
Coronary artery disease N (%)				
No	8469	(88.4)	137	(83.0)
Yes	1108	(11.6)	28	(17.0)
Hypertension N (%)				
No	6141	(64.1)	93	(56.4)
Yes	3436	(35.9)	72	(43.6)
Lab N (%)				
1	4159	(43.4)	67	(40.6)
2	5418	(56.6)	98	(59.4)
Gender N (%)				
Males	7294	(76.2)	132	(80.0)
Females	2283	(23.8)	33	(20.0)
Age N (%)				
<55	1907	(19.9)	33	(20.0)
55-64	2698	(28.2)	61	(37.0)
65-74	2917	(30.5)	47	(28.5)
75	2055	(21.5)	24	(14.5)
Serum N (%)				
1	483	(5.0)	5	(3.0)
2	896	(9.4)	15	(9.1)
3	879	(9.2)	12	(7.3)
4	1214	(12.7)	11	(6.7)
5	743	(7.8)	27	(16.4)
6	1079	(11.3)	30	(18.2)
7	1871	(19.5)	11	(6.7)
8	674	(7.0)	18	(10.9)
9	726	(7.6)	21	(12.7)
10	86	(0.9)	10	(6.1)
11	926	(9.7)	5	(3.0)

uninterpretable tests can be avoided, it may reduce patient anxiety that may result from uncertainty, as well as save costs associated with repeat venipuncture, shipping, and laboratory analysis. To our knowledge, this is the first study to investigate clinical sources related to BeLPT uninterpretable outcomes. We observed that 3.1% of all BeLPT results from two labs over a 10-year period were uninterpretable due to PG or OP. The particular pooled AB-positive serum lot used in the assay stood out as the major factor involved in PG and OP. Additionally, our data show that participant specific factors are not the primary drivers of PG or OP, although several clinical factors for PG, namely hemoglobin concentration, MCHC, and self-reported joint pain, suggest hypotheses for future study. We observed no relationship between other participant specific factors in the multivariable models, for which data were available. Importantly, we did not see evidence of a relationship between clinical risk factors that one might expect to affect this immunoassay, such as cancer and diseases associated with immunosuppression. In addition, there was poor agreement among subject matter experts when they independently and blindly ranked clinical variables that might cause PG and OP. This poor agreement among raters is in line with the notion that potential clinical drivers of uninterpretable results are not well understood.

4.1 | Role of laboratory testing conditions in producing uninterpretable results

BeLPT testing laboratories currently evaluate how the human AB-positive pooled serum that is routinely added to cell culture media impacts lymphocyte proliferation in response to beryllium salts, antigens, and mitogens. Our findings suggest that laboratories should also consider how serum lot variability impacts the occurrence of uninterpretable tests when choosing lots.

Serum lot was found to be important in explaining both PG and OP, accounting for 21% and 16% of the variation, respectively. There was a disproportionate number of uninterpretable results related to a specific serum lots and laboratories (Table 1 and 3). This, coupled with the lack of evidence that participant specific risk factors are driving uninterpretable results, warrants future investigations to examine serum lots and laboratory practices in greater detail. Additionally, we observed that OP thresholds were highly variable across the different serum lots, further reinforcing the importance of serum lot evaluation when assessing lymphocyte proliferation assay performance.

Future studies should examine if adjustment of lymphocyte number in the assay could help optimize the BeLPT. Like all lymphocyte proliferation assays, the BeLPT is dependent on the addition of the patient's white blood cells, including lymphocytes, to the culture medium. Our multivariable analyses showed that lymphocyte count was associated with OP, although it was not significant after accounting for multiple testing. Because the lymphocyte counts in categories other than normal were infrequent, inference on the odds ratios is limited. The classification results showed that the generalized linear mixed model, as well as the CART models, were limited in their ability to distinguish between cases and controls regardless of the amount of oversampling.

TABLE 4 Multivariable model results used to assess the association for Over Proliferation results

Model	OR	95%CI	Test statistic	P-value	FDR adjusted P-value
Over proliferation					
Lymphocyte count			$F_{(2,10)} = 7.44$	0.0105	0.074
Low vs normal	0.328	0.066-1.619			
High vs normal	4.675	1.751-12.481			
Burning pain in hands/feet (yes vs no)	1.392	0.863-2.244	$t_{10} = 0.1542$	0.154	0.154
Numbness/tingling in fingers/toes (yes vs no)	1.366	0.913-2.042	$t_{10} = 0.1149$	0.115	0.154
Coronary artery disease (yes vs no)	1.547	0.93-2.572	$t_{10} = 0.0853$	0.085	0.154
Hypertension (yes vs no)	1.317	0.903-1.92	$t_{10} = 0.135$	0.135	0.154
Gender (female vs male)	0.708	0.451-1.111	$t_{10} = 0.1189$	0.119	0.154
Age			$F_{(3,10)} = 3.66$	0.052	0.154

OR, Odds Ratio; Confidence Interval, CI; FDR, False Discovery Rate.
Results are shown for the model with the lowest AICc. Note: No results were significant after adjusting P-values using the FDR to account for multiple testing. Sample sizes for the model can be found in the participant characteristics overview for Over Proliferation in Table 3.

The limited utility of clinical risk factors in explaining uninterpretable BeLPTs, along with the large amount of variation explained by serum lot, suggest that laboratory/test conditions are likely a driving factor for uninterpretable BeLPTs. However, the assessment of laboratory/test conditions was outside the scope of this study. Our objective was to gain insight into potential participant specific risk factors that may contribute to uninterpretable BeLPTs. While, we controlled for differences in laboratory in our initial analysis, we did not interpret its effect. Upon review this has been noted as an area of interest, in order to identify whether there were any significant differences between laboratories after accounting for the differences in serum lot. Therefore, a post hoc analysis that did not control for the FDR was preformed to assess the effect that laboratory had on uninterpretable results. No significant difference in the effect of laboratory was seen for OP (OR: 1.34, 95%CI: 0.41-4.45, P-value = 0.591). However, a significant difference in the odds of PG was seen between the laboratories (OR: 0.17, 95%CI: 0.04-0.85, P-value = 0.03).

These results suggest that for PG there seem to be differences between the laboratories beyond serum lot that result in increased odds for PG, but not for OP. We are only able to speculate on why differences may exist, but one possible explanation is that there is no standard definition for PG across laboratories. Even though we took great care to standardize the definition based on subject matter expert guidance, it is possible that a test that would have been labeled as PG at one laboratory might be identified differently if it was processed at different laboratory. This would ultimately lead to differences in how a test was entered into the NSSP database and it is possible that they met our exclusion criteria as a result. For example, if a test was entered as a type of invalid test that met out exclusion criteria, that the data was simply never
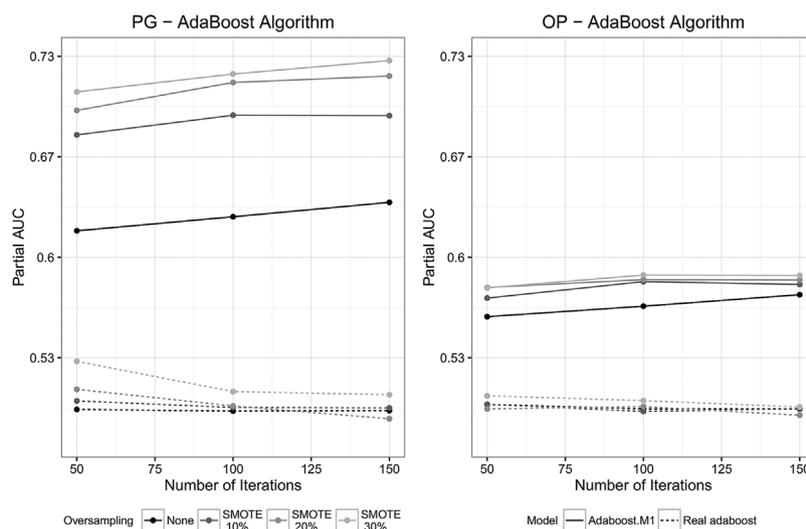
FIGURE 1 CART training results for the AdaBoost Algorithm across various tuning parameters. The final classification model for each degree of oversampling was chosen based on the set of tuning parameters that yielded the highest partial AUC. Prediction was then assessed by applying the final model to the testing dataset
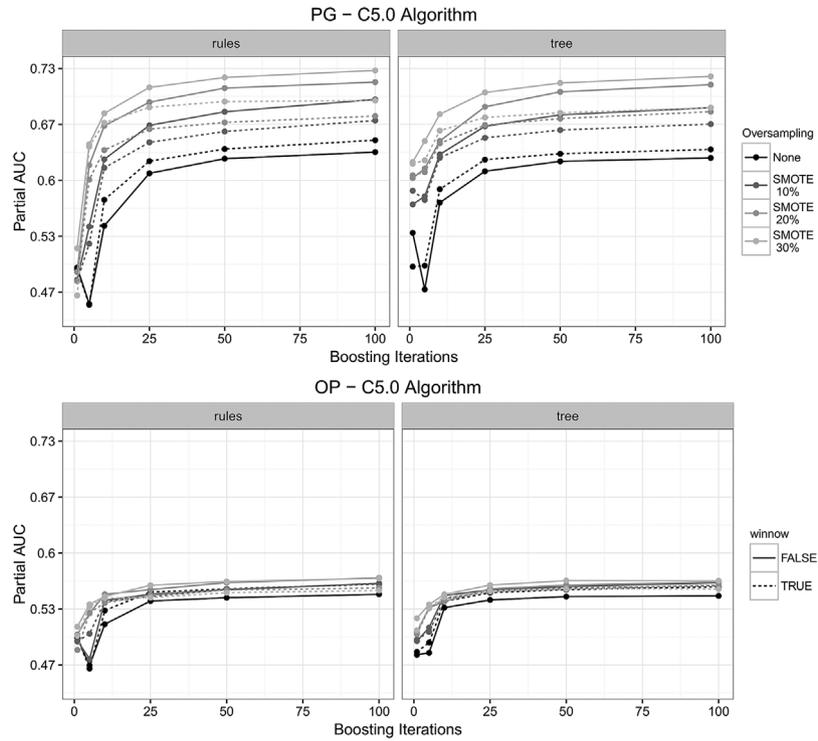
FIGURE 2 CART training results for the C5.0 Algorithm across various tuning parameters. The final classification model for each degree of oversampling was chosen based on the set of tuning parameters that yielded the highest partial AUC. Prediction was then assessed by applying the final model to the testing dataset

captured for us to reassess using the standardized approach we presented here. Additionally, as previously highlighted this could be related differences in laboratory/test conditions, other than serum lot, which are not a part of the data collected through the NSSP. Future insight could be gained by collecting and assessing data on test/laboratory conditions, such as age of serum, sample collection information, and shipping conditions.

4.2 | Role of clinical characteristics in producing uninterpretable results

Numerous variables displayed a strong association with uninterpretable results in the univariate analysis, however many of these associations did not persist in the multivariable setting models. Differences in the proportion of uninterpretable results related to a given laboratory and
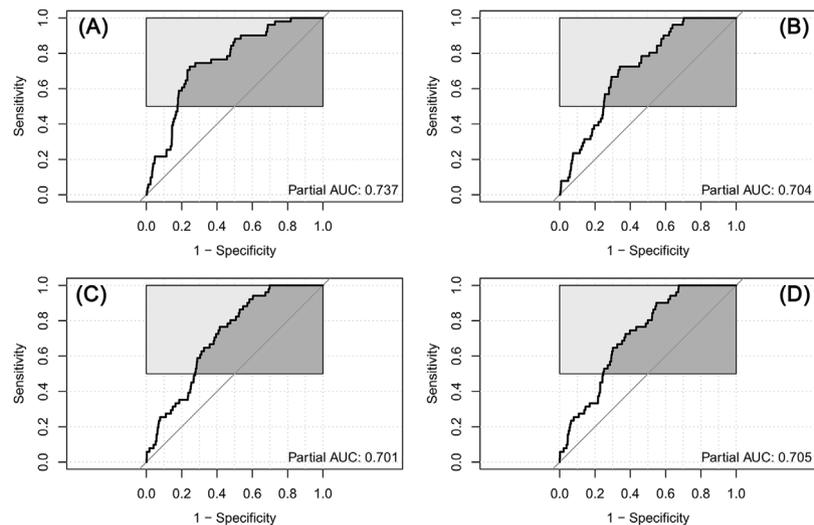
FIGURE 3 ROC results for the PG and the multivariable generalized linear mixed model across varying degrees of oversampling. Panel (A) no oversampling, (B) 10% oversampling, (C) 20% oversampling, and (D) 30% oversampling

TABLE 5 Partial AUC results for the classification models

Model	None [Mean (95%CI)]	SMOTE 10% [Mean (95%CI)]	SMOTE 20% [Mean (95%CI)]	SMOTE 30% [Mean (95%CI)]
Poor growth				
AdaBoost	0.690 (0.688-0.692)	0.704 (0.701-0.706)	0.705 (0.702-0.707)	0.702 (0.700-0.705)
C5.0	0.50 (0.50-0.50)	0.712 (0.710-0.714)	0.736 (0.734-0.738)	0.710 (0.708-0.711)
Mixed model	0.737 (0.735-0.739)	0.704 (0.703-0.706)	0.702 (0.701-0.704)	0.707 (0.705-0.709)
Over proliferation				
AdaBoost	0.612 (0.610-0.614)	0.630 (0.628-0.633)	0.652 (0.650-0.654)	0.590 (0.588-0.592)
C5.0	0.50 (0.50-0.50)	0.619 (0.617-0.620)	0.608(0.606-0.610)	0.638 (0.636-0.640)
Mixed model	0.663 (0.661-0.665)	0.633 (0.631-0.635)	0.622 (0.620-0.624)	0.590 (0.588-0.592)

AUC, Area Under the Curve; CI, Confidence Interval; SMOTE, Synthetic Minority Over-sampling Technique.
Mean and 95% z-based CI estimates were determined using bootstrapping with 1000 replicates.

serum lot further reinforce the need to control for these variables when evaluating participant specific risk factors (Table 1 and 2). PG displayed a significant association with hemoglobin concentration, MCHC, and joint pain. Both hemoglobin concentration and MCHC proved to be protective against PG in univariate analyses. We speculate that either the hemoglobin molecule itself or the iron that is bound by hemoglobin are beneficial nutrients in short term lymphocyte culture. Culture media, specifically RPMI 1640, does not include either hemoglobin or iron. Future laboratory studies are warranted to examine how addition of hemoglobin and/or iron to the culture medium impacts proliferation, especially in light of the literature showing that iron affects human lymphocyte proliferative responses.[34–36] Alternatively, we have no explanation for the observation that joint pain was found to be protective of PG. This might be an artifact of residual confounding.

Cancer treatment can also result in reduced lymphocyte growth.[37] Yet, the numerous individual cancer-related variables in the NSSP dataset (eg, histories of cancer in the lung, colon, breast, stomach, and skin) did not show a strong enough association with PG to be included in the final multivariable model. In addition, some of these cancer variables were not explored due to the low prevalence in the study population. A composite cancer variable indicating if one had ever had lung, colon, breast, stomach, and/or skin was created to try to evaluate cancer more holistically, but was not a strong enough univariate factor to enter the final multivariable model. Additionally, diabetes mellitus may impair lymphocyte function.[38] However, blood glucose levels and history of diabetes did not display a strong enough association with PG in the univariate setting to be included in the final multivariable model. It has been hypothesized that age may also influence lymphocyte growth. For example, older women have shown significantly lower lymphocyte proliferation compared to younger women.[39] While, the percentage of those with PG increased by age group and is consistent with these findings, we observed no significant relationship between PG and age. Similarly, we did not find a significant relationship between gender and PG. However because our cohort tended to be older and predominately male compared to the U.S. population (2010 Census), generalizability to the general population is limited.

An important limitation is that many of the clinical variables are self-reported and subject to recall bias. In addition, the rare event nature of both the outcomes (ie, PG and OP) as well the rare occurrence of many of the health conditions in the dataset (eg, cancer, CBD, rheumatoid arthritis, etc.) limited our assessment of some potentially important clinical risk factors. Missing records were present for numerous individuals, which is believed to be attributed to the algorithms used by the NSSP to select clinical tests based on an individual's past occupational exposures. Thus, we assumed that these records were missing at random and did not attempt to impute these values when evaluating their association with the outcomes. As a result the cohort used in the multivariable models when assessing the association was smaller (PG n = 115, OP n = 165), since individuals were excluded if they did not have complete records for the variables of interest (Table 1 and 3). Lastly, when evaluating the association between patient characteristics and the outcomes, we attempted to assess and select the set of variables that explained the most amount of variation in the outcome of interest. Yet, we were not able to evaluate all possible combinations and their interactions. There may have been other personal health conditions and certain medications that could have contributed to PG or OP that we were unable to study in the available dataset. For example, some personal health conditions that have been shown to influence lymphocyte growth include exercise, performance enhancing drugs, vitamins, and the human rhinovirus.[40–45] Nonetheless, the NSSP dataset is one of the most comprehensive collections of lymphocyte proliferation assay data coupled with medical examination data. As such, it currently provides the best grounds on which to evaluate the association between participant specific factors and uninterpretable BeLPTs due to PG and OP.

5 | CONCLUSIONS

Uninterpretable BeLPTs due to PG and OP were found to be related to the selection of serum lot for the assay and were not shown to be driven by participant specific factors. There was evidence that PG may be influenced by hemoglobin and/or iron. Continued insight into uninterpretable BeLPTs would benefit from a standardization of practices and definitions across laboratories. The role of hemoglobin and lymphocyte count merit consideration when attempting to drive

down the prevalence of uninterpretable BeLPTs. The lack of clinical risk factors as drivers of uninterpretable results highlights the need for future studies to focus on laboratory/test conditions as potential drivers of uninterpretable results.

AUTHORS' CONTRIBUTIONS

DES participated in the conception and design of work, analyses, and interpretation. Drafted and revised work. Participated in final approval of work and agree to be accountable for all aspects of work. APG participated in the conception and design of work, data acquisition, analyses, and interpretation. Critically reviewed work for intellectual content. Participated in final approval of work and agree to be accountable for all aspects of work. AWS participated in the conception and design of the work, acquisition and interpretation of the data, and revisions of the draft. Agree to be accountable for all aspects of work. EB participated in the conception and design of work, acquisition and analysis of the data for the work, and revision of the draft. Agree to be accountable for all aspects of the work. MM participated in the conception and design of work, acquisition and analysis of the data for the work, and revision of the draft. Agree to be accountable for all aspects of the work. AEB participated in analysis planning, review and revisions of the work, and final approval. DG participated in analysis planning, review and revisions of the work, and final approval. LM participated in the conception and design of work, reviewed analyses, and discussed interpretation of the analyses. DC participated in the conception and design of work, reviewed analyses, discussed interpretation of the analyses, reviewed and commented on the manuscript, participated in the final approval of the work, and agree to be accountable for all aspects of work. LSN participated in the conception and design of the NSSP surveillance program, the work, analyses, and interpretation of the study findings. Provided clinical and laboratory expertise regarding the Be-LPT. Participated in preparation of manuscript, final approval of work and agree to be accountable for all aspects of the work.

ETHICS APPROVAL AND INFORMED CONSENT

IRB approval with exempt status was obtained through the Central Department of Energy Institutional Review Board (FWA 00015 567).

DISCLOSURE (AUTHORS)

The authors declare no conflicts of interest.

DISCLOSURE BY AJIM EDITOR OF RECORD

Steven B. Markowitz declares that he has no competing or conflicts of interest in the review and publication decision regarding this article.

DISCLAIMER

None.

ORCID

Derek E. Smith http://orcid.org/0000-0001-8117-2477
Ashley P. Golden http://orcid.org/0000-0002-9430-7702

REFERENCES

1. Newman LS, Mroz MM, Balkissoon R, Maier LA. Beryllium sensitization progresses to chronic beryllium disease: a longitudinal study of disease risk. *Am J Respir Crit Care Med*. 2005;171:54–60.
2. Kreiss K, Day GA, Schuler CR. Beryllium: a modern industrial hazard. *Annu Rev Public Health*. 2007;28:259–277.
3. Newman L. Significance of the blood beryllium lymphocyte proliferation test. *Environ Health Perspect*. 1996;104:953–956.
4. Stange AW, Furman FJ, Hilmas DE. The beryllium lymphocyte proliferation test: relevant issues in beryllium health surveillance. *Am J Ind Med*. 2004;46:453–462.
5. Middleton DC, Mayer AS, Lewin MD, Mroz MM, Maier LA. Interpreting borderline BeLPT results. *Am J Ind Med*. 2011;54:205–209.
6. Middleton DC, Fink J, Kowalski PJ, Lewin MD, Sinks T. Optimizing BeLPT criteria for beryllium sensitization. *Am J Ind Med*. 2008;51:166–172.
7. Frome EL, Newman LS, Cragle DL, Colyer SP, Wambach PF. Identification of an abnormal beryllium lymphocyte proliferation test. *Toxicology*. 2003;183:39–56.
8. DOE 2001. DOE specification beryllium lymphocyte proliferation testing (BeLPT). U.S. Department of Energy Washington D.C. DOE-SPEC-1142–2001.
9. OSHA Occupational exposure to beryllium. *Fed Regist*. 2017; 82:2470–2757.
10. Hines SE, Pacheco K, Maier LA. The role of lymphocyte proliferation tests in assessing occupational sensitization and disease. *Curr Opin Allergy Clin Immunol*. 2012;12:102–110.
11. Kwon YM, Thomas P, Summer B, et al. Lymphocyte proliferation responses in patients with pseudotumors following metal-on-metal hip resurfacing arthroplasty. *J Orthop Res*. 2010;28:444–450.
12. Hoffman KM, Ho DG, Sampson HA. Evaluation of the usefulness of lymphocyte proliferation assays in the diagnosis of allergy to cow's milk. *J Allergy Clin Immunol*. 1997;99:360–366.
13. Cederbrank K, Hultman P, Marcusson AJ, Tubbling L. In vitro lymphocyte proliferation as compared to patch test using gold, palladium and nickel. *Int Arch Allergy Immunol*. 1997;112:212–217.
14. De Vuyst P, Dumortier P, Schandené L, Estenne M, Verhest A, Yernault JC. Sarcoidlike lung granulomatosis induced by aluminum dusts. *Am Rev Respi. Dis*. 1987;135:493–497.
15. Stange B, McInerney J, Golden A, et al. Integrated approach to health screening of former department of energy workers detects both

occupational and non-occupational illness. *Am J Ind Med.* 2016;59:200–211.

16. United States General Accounting Office (GAO). 2001. Government responses to beryllium uses and risks. Pub. No. GAO/OCG-00-6. Washington, D.C.

17. Henneberger PK, Goe SK, Miller WE, Doney B, Groce DW. Industries in the United States with airborne beryllium exposure and estimates of the number of current workers potentially exposed. *J Occup Environ Hyg.* 2004;1(10):648–659.

18. SAS Institute 2012. The SAS system for Windows. Release 9.4. SAS Inst Cary, NC.

19. Bolker B, Bates D, Maechler M, et al. 2015. lme4: Linear mixed-effects models using 'eigen' and S4. R package version 1.1-10.

20. R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org [Accessed 06/05/2017].

21. Krippendorff K. 2011. Computing Krippendorff's alpha-reliability. http://repository.upenn.edu/asc_papers/43. Accessed June 5, 2017.

22. Calcagno V. 2013. glmulti: Model selection and multimodel inference made easy. R package version 1.0. 7.

23. Benjamini Y, Hochbery Y. Controlling the false discovery rate: a practical and powerful method approach to multiple testing. *J R Statist Soc B.* 1995;57:289–300.

24. Dobbin K, Simon R. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Med Genomics.* 2011;4:31.

25. Torgo L. 2013. DMwR: Functions and data for "data mining with R." R package version 0.4. 1.

26. Kang J, Schafer L. 2010. Estimating average treatment effected when the treatment is a latent class. http://stat.psu.edu/research/technical-report-files/2010-technical-reports/TR10-05.pdf. Accessed July 18, 2017.

27. Kuhn M, Johnson K. 2013. *Regression trees and rule-based models. In: Applied predictive modeling.* New York: Springer. pp. 173–218.

28. Kuhn M. 2016. caret: Classification and Regression Training. R package version 6. 0–73.

29. Xavier R, Natacha T, Alexandre H, et al. 2017. pROC: Display and Analyze ROC Curves. R package version 1.10. 0.

30. McChlish DK. Analyzing a portion of the ROC curve. *Med. Decis. Making* 1989;9:190–195.

31. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Statist* 2000;28:337–407.

32. King G, Zeng L. Logistic Regression in rare events data. *Political Analysis* 2001;9:137–163.

33. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artificial Intelligence Res.* 2002;16:321–357.

34. Kaur S, Srivastava G, Sharma AN, Jolly RS. Novel immunosuppressive agent caerulomycin A exerts its effect by depleting cellular iron content. *Br J Pharmacol.* 2015;172:2286–2299.

35. Zaczyńska E, Kocięba M, Śliwińska E, Zimecki M. Bovine lactoferrin enhances proliferation of human peripheral blood lymphocytes and induces cytokine production in whole blood cultures. *Adv Clin Exp Med.* 2014;23:871–876.

36. Seligman PA, Kovar J, Gelfand EW. Lymphocyte proliferation is controlled by both iron availability and regulation of iron uptake pathways. *Pathobiology* 1992;60:19–26.

37. Barsegian V, Hueben C, Mueller SP, et al. Impairment of lymphocyte function following yttrium-90 DOTATOC therapy. *Cancer Immunol Immunother.* 2015;64:755–764.

38. Otton R, Mendonça JR, Curi R. Diabetes causes marked changes in lymphocyte metabolism. *J Endocrinol.* 2002;174:55–61.

39. Meydani SN, Endres S, Woods MM, et al. Oral (n-3) fatty acid supplementation suppresses cytokine production and lymphocyte proliferation: comparison between young and older women. *J Nutr.* 1991;121:547–555.

40. Aab A, Wirz O, van de Veen W, et al. Human rhinoviruses enter and induce proliferation of B lymphocytes. *Allergy.* 2017;72:232–243.

41. Arakawa K, Hosono A, Shibata K, et al. Changes in blood biochemical markers before, during, and after a 2-day ultramarathon. *J Sports Med.* 2016;7:43–50.

42. Buonocore D, Negro M, Arcelli E, Marzatico F. Anti-inflammatory dietary interventions and supplements to improve performance during athletic training. *J Am Coll Nutr.* 2015;34:62–67.

43. Imperlini E, Spaziani S, Mancini A, Caterino M, Buono P, Orrù S. Synergistic effect of DHT and IGF-1 hyperstimulation in human peripheral blood lymphocytes. *Proteomics* 2015;15:1813–1818.

44. Moreira A, Kekkonen RA, Delgado L, Fonseca J, Korpela R, Haahtela T. Nutritional modulation of exercise-induced immunodepression in athletes: a systematic review and meta-analysis. *Eur J Clin. Nutr.* 2007;61:443–460.

45. Dohi K, Mastro AM, Miles MP, et al. Lymphocyte proliferation in response to acute heavy resistance exercise in women: influence of muscle strength and total work. *Eur J Appl Physiol.* 2001;85:367–373.