PharmaSUG 2021 - Paper RW-209

Standardizing Laboratory Results from Diverse Real-World Data to Enable Meaningful Assessments of Drug Safety and Effectiveness

Irene Cosmatos and Michael Bulgrien, United BioSource LLC

ABSTRACT

The highly granular clinical details in electronic medical record (EMR) data are increasingly critical for active safety surveillance of marketed drugs. Many of the Food and Drug Administration's (FDA) Sentinel's drug safety queries were unsuccessful because administrative claims data were 'insufficient' due to lack of laboratory results.

The study's objective was to standardize laboratory data from the United States (US) and non-US EMR databases into a common structure to enable active safety surveillance analyses to be performed and compared across data sources that do not use a standardized coding system such as Logical Observation Identifiers, Names, and Codes (LOINC).

UBC's database analysts and clinicians developed an approach to transform source data into a cohesive and accurate dataset of laboratory results based on standardized units and test names, while minimizing loss of data. Steps included 1) performing an algorithmic search using keywords from the FDA's Sentinel test definitions to select appropriate laboratory names; 2) review and acceptance by European (EU) and US clinicians of test names, test specimen type, and units; and 3) verification of unit matching or conversion of units to a single valid unit for each test type. The initial effort focused on 3 liver function tests (LFTs) that are important for safety assessment: alanine aminotransferase (ALT), aspartate transaminase (AST, also known as SGPT), and total bilirubin. Three real-world data sources were included: 2 US and 1 EU.

Across the three databases, the algorithmic search discovered 107 unique test names that represented LFTs (Step 1). However, slightly more than half of these initial matches (59, 55%) were excluded from the final LFT group after clinician review (Step 2), due to incorrect initial classification by the algorithm (23%); test name indicated a different clinical measurement, such as '*direct* bilirubin' rather than '*total* bilirubin' (18%); non-acceptable specimen type (13%); or other reasons (0.9%). The 48 clinically approved LFT names were associated with 149 unique units (e.g., total bilirubin measured in mmol/L and mg/dL). After removing units that could not be matched or converted to the standard (Step 3), our transformed LFT database contained 260 million records. Importantly, only 1.1% of the original LFT records could not be included as a valid LFT result.

Diverse and 'messy' laboratory data found in real-world EMR datasets can be successfully converted into a standardized structure for meaningful safety evaluations. However, close clinical scrutiny and unit conversions are required before a common data structure is available.

INTRODUCTION

Laboratory results available in EMR data are becoming increasingly critical for the assessment of drug safety and effectiveness. However, unlike diagnoses and medications where vocabularies for data capture are well established, the recording of laboratory results in EMR databases is often not standardized, with differences observed across and even within EMR systems. One proposed standard laboratory coding system, LOINC has thus far had limited

adoption in healthcare EMR systems due to several factors, including its own complexity (e.g., there are 60+ different LOINC codes for testing blood glucose levels).

The heterogeneity in recording of laboratory data creates significant challenges for epidemiologic analyses requiring cohort and outcome definitions based on laboratory criteria. This paper describes UBC's approach to maximize the value of laboratory data in diverse EMR databases through a standardization process involving all elements of laboratory data: name of test, specimen used, test conditions (e.g., fasting), test results (quantitative and qualitative), and units of test results. UBC's database analysts, working closely with US and EU clinicians familiar with laboratory data, developed an approach to transform laboratory results from diverse EMR databases into a common data structure while minimizing loss of data.

METHODS

OVERVIEW OF UBC'S LABORATORY DATA STANDARDIZATION PROCESS

UBC's database analysts and clinicians developed a 3-step approach to transform source EMR laboratory data into a cohesive and accurate common data structure. The focus of the research was to ensure the resulting common data structure accurately grouped all relevant test names, per data source, for each of the Sentinel's Initiative list of 29 laboratory tests (FDA Sentinel Initiative, 2015) that are important for safety assessment among a drug-exposed population. The Sentinel rules for standardization of test names and results were used as a reference whenever possible. Test result units were converted wherever necessary, using Sentinel's conversion formulas, to the standard unit identified in the Sentinel's reference list (e.g., mg/DL for blood glucose levels). Laboratory results that were successfully placed into one of the 29 test categories but either (1) did not have units or (2) the associated units could not be converted to the standard were retained in the final common dataset in a separate table.

The three steps for the standardization process are described below and in Figure 1.

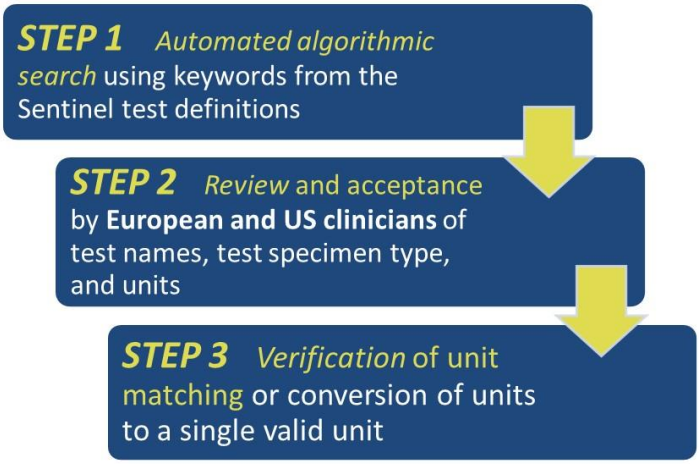

Figure 1. Three-step approach to standardize diverse laboratory data from EMR databases

<u>STEP 1.</u> Perform an algorithmic search using text strings that contain relevant keywords from the FDA's Sentinel test definitions to select appropriate laboratory names.

A ‘preliminary’ list of relevant exclusion criteria for each laboratory name was created during this step based on Sentinel keywords. The lists were then updated in Step 2, after clinical review, to prepare the final set of inclusion and exclusion criteria per database.

An example of included and excluded text strings that were used to identify appropriate laboratory names for a pregnancy test is provide in Table 1.

Search String	Inclusion/Exclusion Criteria
down synd	Exclude
monochorio	Exclude
multiple	Exclude
sti*	Exclude
gonadotropin	Include
hcg	Include
pregnancy	Include

Table 1. Inclusion and exclusion text string search criteria for identifying a pregnancy test

The *exclusion* criteria removed pregnancy-related laboratory names that contained the term ‘pregnancy’ but were not specific to determining pregnancy status, such as:

- PREV PREGNANCY DOWN SYND
- PREGNANCY MONOCHORIONIC
- MULTIPLE PREGNANCY
- STI PREGNANCY PANEL W/O PAP
- STI PREGNANCY PANEL W/PAP
- STI SCREEN/PREGNANCY W/O PAP
- STI SCREEN/PREGNANCY W/PAP

The *inclusion* criteria added laboratory names that were relevant for testing if a woman was pregnant even if the names did not contain ‘pregnancy’, such as:

- BETA SUBUNIT, HCG, SERUM
- BETA-HCG/M&F
- CHORIOGONADOTROPIN
- CHORIOGONADOTROPIN.BETA SUBUNIT
- CHORIONIC GONADOTROPIN
- CHORIONIC GONADOTROPIN, TOTAL
- CHORIONIC GONADOTROPIN-HCG
- GONADOTROPIN
- GONADOTROPIN RELEASING HORMONE
- HCG
- HCG BETA SUB
- HCG EIA VENTRESCREEN
- HCG PREGNANCY
- HCG SERUM
- HCG TITRE – URINE
- HIGH SENSITIVITY URINE PREGNANCY TEST
- HUMAN CHORIONIC GONADOTROPIN
- HUMAN CHORIONIC GONADOTROPIN (HCG)

- PREGNANCY TEST
- PREGNANCY TEST (URINE)
- PREGNANCY TEST SERUM
- PREGNANCY TEST URINE
- PREGNANCY TEST, SERUM
- PREGNANCY TEST, SR
- PREGNANCY TEST, URINE
- PREGNANCY TEST, URINE (GTSC)
- SERUM PREGNANCY TEST
- SERUM PREGNANCY TEST (B-HCG)
- SERUM TOTAL HCG LEVEL
- TOTAL HCG LEVEL
- URINE HCG
- URINE HCG TITRE
- URINE PREGNANCY
- URINE PREGNANCY TEST

STEP 2. Review and acceptance by EU and US clinicians of the results that were generated from the automated algorithmic search in Step 1.

A clinician representing the country of origin of each database involved in the research effort was required, due to occasional differences by country in the naming conventions of laboratory tests (e.g., 'Hemoglobin A1C' in the US may be reported as 'Haemoglobin A1c' or 'HbA1c' in the EU and 'Prothrombin.INR' in the US may be reported as 'International normalised ratio' in the EU).

The clinicians reviewed every laboratory record that was a successful match to the Step 1 criteria to determine if the laboratory record should be considered a 'true match' based on their clinical judgement. Laboratory names not considered by the clinicians to be an acceptable match were removed via exclusion criteria as demonstrated in the "Pregnancy Test" example above.

STEP 3. Verification of unit matching and conversion to a single valid unit for each test type based on conversion formulas from the FDA Sentinel Initiative.

Unit inclusion criteria identified individual laboratory values that were converted to the Sentinel-specified standard unit using the formula:

$$\text{ConvertedValue} = \frac{\text{OriginalValue} + \text{ConversionAddBefore}}{\text{ConversionDenominator}} + \text{ConversionAddAfter}$$

OriginalValue = The original data-provided lab measurement in a non-standardized unit.
ConversionAddBefore = A value added ***before*** the conversion divisor was applied.
ConversionDenominator = The conversion divisor applied to the original value.
ConversionAddAfter = A value ***after*** the conversion divisor was applied.
ConvertedValue = The converted lab measurement in a specified standardized unit.

Parameters required to transform the original source laboratory value into the standardized unit were provided in the Sentinel common data model laboratory result table documentation.

Consider, for example, the conversion of total bilirubin laboratory values measured in micromoles per liter (μmol/l) to standardized lab units of miligrams per deciliter (mg/dl):

mg/dl = μmol/l × 0.0585

This conversion formula provides a conversion *multiplier* (0.0585). The *divisor* (ConversionDenominator) in the formula above is derived by expressing the conversion *multiplier* (*n*) as a fraction (*n*/1) and inverting it into its reciprocal (1/*n*) using the Inverse Property of Multiplication (Montis 2010):

Any given number multiplied by its reciprocal is equal to one.

$n \times 1/n = 1$	*inverse property of multiplication*
$n = 1 \div 1/n$	divide both sides by the reciprocal (1/*n*)
$1 \times n = 1 \div 1/n$	multiplying by *n* is the same as dividing by its reciprocal (1/*n*)

In other words, any conversion multiplier can be inverted into its reciprocal to be used as a conversion divisor:

mg/dl = μmol/l ÷ (1 / 0.0585) = μmol/l ÷ 17.09401709

A small percentage of bilirubin laboratory values were expressed in other mole-based units. The various mole-based units were converted to micromoles prior to conversion to miligrams per deciliter.

Conversions from Moles	Conversions to Micromoles
1 mole = 1 mol *(mole)*	1 mol *(mole)* = 0.000001 μmol
1 mole = 1,000 mmol *(millimoles)*	1 mmol *(millimole)* = 0.001 μmol
1 mole = 1,000,000 μmol *(micromoles)*	1 umol *(micromole)* = 1 μmol
1 mole = 1,000,000,000 nmol *(nanomoles)*	1 nmole *(nanomole)* = 1,000 μmol

Table 2. Mole conversions

The micromole conversions were integrated with the unit conversion for each laboratory unit variation as follows:

Conversion to Standardized Unit *(expressed as a conversion multiplier)*	Conversion to Standardized Unit *(expressed as a conversion divisor)*
mg/dl = mol/l × 0.0585 / 0.000001	mg/dl = mol/l ÷ (0.000001 / 0.0585)

mg/dl = mmol/l × 0.0585 / 0.001	mg/dl = mmol/l ÷ (0.001 / 0.0585)
mg/dl = umol/l × 0.0585 / 1	mg/dl = umol/l ÷ (1 / 0.0585)
mg/dl = nmole × 0.0585 / 1,000	mg/dl = nmole/l ÷ (1,000 / 0.0585)

Table 3. Moles per liter conversions to miligrams per deciliter

Using the reciprocal of the conversion multiplier, the mole conversion factor was divided by the standardized unit conversion factor (0.0585) to populate the ConversionDenominator parameter in the conversion lookup table for each originating bilirubin total unit type:

Originating Lab Unit	Standardized Lab Unit	Conversion AddBefore	ConversionDenominator	Conversion AddAfter
mg/dl	mg/dl	0.0	1.0	0.0
mol/l	mg/dl	0.0	0.0000170940170940171	0.0
mmol/l	mg/dl	0.0	0.0170940170940171	0.0
umol/l	mg/dl	0.0	17.0940170940171	0.0
nmol/l	mg/dl	0.0	17,094.0170940171	0.0

Table 4. Lab results conversion lookup table for total bilirubin

In practice, clinical review may uncover a scenario in which laboratory values for one or more unit variations do not appear to be accurately represented in the source data. For example, one would expect percentiles for different mole-based units to differ by a factor of 1,000 or more. When this is not the case, the accuracy of those laboratory values is called into question and the corresponding units should be excluded from the conversion into standardized laboratory units.

For some lab unit conversions, a numeric offset must be applied before or after the lab unit conversion factor is applied. According to the FDA Sentinel, glycated hemoglobin (HbA1c) laboratory measurements are one example of this:

> "For any records with an Orig_Result_unit of "mmol/mol" (or any variation of this), MS_Result_N is converted to a percentage using the following equation: HGBA1c % = (Orig_Result/10.929) + 2.15"

Therefore, the conversion parameter lookup table for HbA1c contained the following row for measurements expressed in micromoles per mole unit values:

Originating Lab Unit	Standard Lab Unit	Conversion AddBefore	ConversionDenominator	Conversion AddAfter
mmol/mol	PERCENT	0.0	10.929	2.15

BODY MASS INDEX (BMI) COMPUTATION USING LOINC

Not all laboratory or body measurements were converted into standardized units. In cases where these values remained in their originating laboratory units, measurement conversions

were done dynamically during the data analysis. One example was BMI computations using LOINC. When BMI observations were not available in a patient's EMR, BMI was calculated using dynamic conversion of weight and height observations into standardized units (kilograms and meters) for use in the equation BMI = kg/m^2.

Category	Units	LOINC Code	LOINC Description
BMI	kg/m^2	39156-5	Body mass index (BMI) [Ratio]
BMI		89270-3	Body mass index (BMI) [Ratio] Estimated
Weight	Standard Unit = kg *Unit Conversions:* oz × 0.0283495 lb × 0.453592 g × 0.001 kg × 1.0	29463-7	Body weight
Weight		75292-3	Body weight - Reported --usual
Weight		79348-9	Body weight --used for drug calculation
Weight		8335-2	Body weight Estimated
Weight		3141-9	Body weight Measured
Weight		8344-4	Body weight Measured --post dialysis
Weight		8348-5	Body weight Measured --pre pregnancy
Height	Standard Unit = m *Unit Conversions:* in × 0.0254 ft × 0.3048 cm × 0.01 m × 1.0	8302-2	Body height
Height		8306-3	Body height --lying
Height		91370-7	Body height --used for drug calculation
Height		8301-4	Body height Estimated
Height		3137-7	Body height Measured
Height		3138-5	Body height Stated

Table 5. LONIC codes and descriptions for BMI

ALT, AST, AND TOTAL BILIRUBIN

To illustrate the details of the standardization process described above, this paper presents the results of the process for a subset of the 29 Sentinel laboratory tests. Three LFTs that are important for the safety evaluation of drug exposure were chosen: ALT, AST or SGPT, and total bilirubin.

DATA SOURCES

Laboratory data from 3 real-world data sources (2 US and 1 EU) were used to illustrate the standardization process.

The Premier Healthcare Database

The Premier Healthcare Database (PHD) is a large US hospital-based, service-level, all-payer database containing discharge information from inpatient and hospital-based outpatient visits. It represents approximately one-quarter of all US admissions from geographically diverse non-governmental community and teaching hospitals and rural and urban health systems. The PHD contains data from standard hospital discharge files, including patient demographics and disease states; health insurance type; admission and discharge diagnoses; admission source and type; discharge status and disposition; and hospital pharmacy medication use. Unique

masked identifiers allow patient data to be followed within the same hospital and across inpatient and hospital-based outpatient settings. All data in the PHD are statistically de-identified.

Optum Humedica Inc.

The EHR data contain information for more than 70 million patients in the US, of whom almost two-thirds are associated with integrated delivery networks (IDN), that provide a spectrum of healthcare services. These data are clinically rich and include laboratory results; vital signs; body measurements; lifestyle observations; biomarkers; inpatient and outpatient treatments, including written prescriptions; inpatient-administered medications; and provider notes. Diagnoses and procedures performed are coded using ICD-9-CM and ICD-10-CM. Medication use is coded by the National Drug Codes (NDC) and Healthcare Common Procedure Coding System (HCPCS), J codes.

Clinical Practice Research Datalink (CPRD)

CPRD is a UK EMR database that is jointly sponsored by the Medicines and Healthcare Products Regulatory Agency and the National Institute for Health Research (NIHR). The database represents de-identified patient data from a network of general practitioner (GP) practices across the UK. Primary care data are linked to other health related data to provide a longitudinal, representative UK health dataset (e.g., linkage to Hospital Episode Statistics (HES)). There are 60 million patient lives, including 12 million currently registered. For >30 years, research has informed clinical guidance and best practice, resulting in over 2,700 peer-reviewed publications.

RESULTS

Across the three databases, the algorithmic search identified 107 unique test names that represented *potential* matches to one of the LFTs (Step 1). However, slightly more than half of these initial matches (59, 55%) were excluded from the final set of acceptable LFT test names after clinician review (Step 2), due to incorrect initial classification by the algorithm (23%); test name indicated a different clinical measurement, such as ‘direct bilirubin’ rather than ‘total bilirubin’ (18%); non-acceptable specimen type (13%); or other reasons (0.9%). The final count of clinically approved unique LFT names after Step 2 was 41. Table 6 lists the final ALT and total bilirubin source names approved by the clinicians. Figure 2 presents the reasons for excluding a laboratory name that had initially been identified in Step 1 from the final list after clinical review.

Lab Type	Source Lab Name
ALT	Alanine aminotransferase (ALT)
	alanine aminotransferase (SGPT), serum
	alanine aminotransferase (SGPT), serum, outside laboratory
	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma
	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma by Without P-5'-P
	alanine, plasma, quantitative
	ALT/SGPT level abnormal
	ALT/SGPT level normal
	ALT/SGPT serum level
	Plasma alanine aminotransferase level
	Serum alanine aminotransferase level
Bilirubin	Bilirubin in sample
	Bilirubin level
	Bilirubin profile
	bilirubin, body fluid
	bilirubin, neonatal, total
	bilirubin, serum, maximum
	bilirubin, serum, total
	BILIRUBIN, TOTAL
	Bilirubin.neonatal.total
	Bilirubin.total
	Bilirubin.total [Mass/volume] in Serum or Plasma
	Bilirubin. Transcutaneous
	Fluid sample bilirubin
	Plasma total bilirubin level
	Serum bilirubin borderline
	Serum bilirubin level
	Serum bilirubin normal
	Serum bilirubin NOS
	Serum bilirubin raised
	Serum total bilirubin level
	Total bilirubin
	Transcutaneous bilirubin

11 Clinically acceptable matches for ALT (SGPT)

22 Clinically acceptable matches for bilirubin

Table 6. ALT and bilirubin source names approved by clinicians

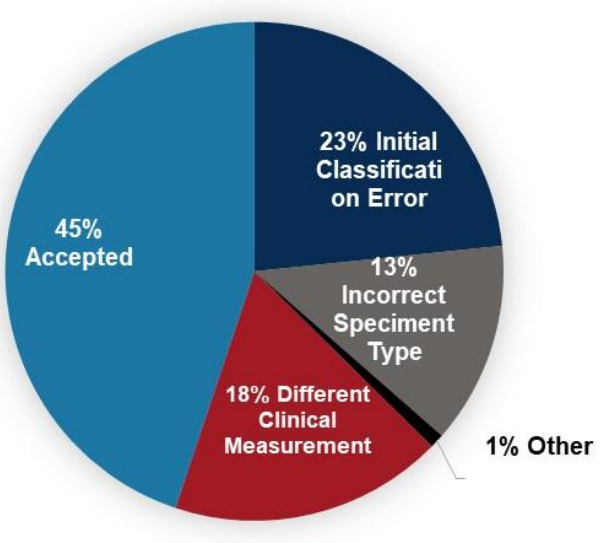

Figure 2. Reasons for exclusions from final LFT file

The 48 clinically approved LFT names were associated with 149 unique units (e.g., total bilirubin measured in mmol/L and mg/dL). After removing units that could not be matched or converted to the standard (Step 3), the transformed LFT database contained 260 million records. Importantly, overall, only 1.1% of all original LFT records could not be included in our standardized laboratory dataset as a valid LFT result. (Figure 3)

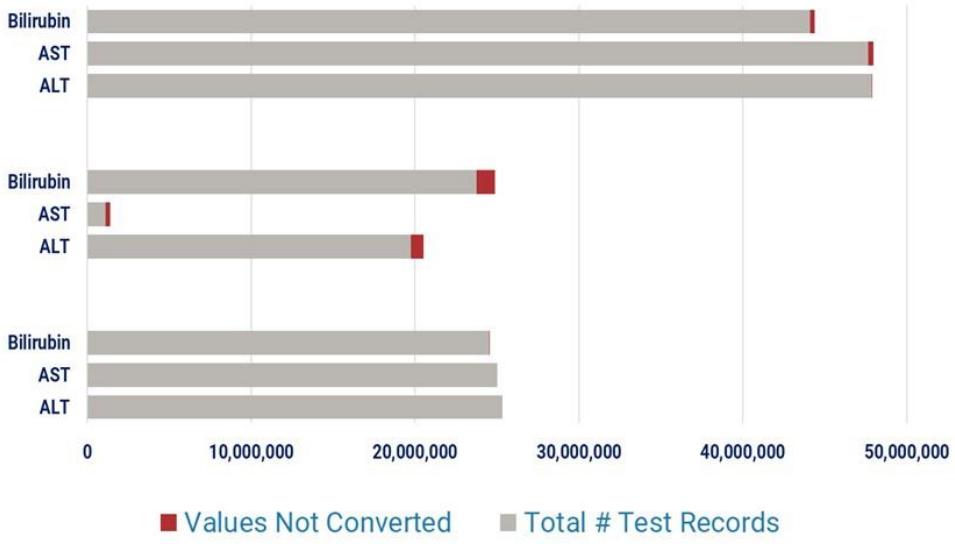

Figure 3. Final results for conversion of laboratory data to standard test names and common units

CONVERSION OF THE LOWER LIMIT OF NORMAL (LLN) AND UPPER LIMIT OF NORMAL (ULN)

Laboratory test data generally included the LLN and ULN associated with the laboratory that conducted the test, which were needed to interpret the clinical relevance of a quantitative laboratory result. Whenever conversion of the units associated with a test result was successful,

the same conversion was applied to the ULN and ULN values to maintain their computational integrity.

CONCLUSION

Research-ready EMR data have become increasingly available in recent years, including valuable laboratory test results that reflect patient's historical laboratory data as well as changes in laboratory values going forward in time as patients initiate new medications to treat underlying diseases. Challenges arise, however, due to the lack of standardization and overall 'dirtiness' of the laboratory data captured within and across EMR systems, often preventing meaningful evidence to be generated in an efficient, timely manner.

This paper introduces a semi-automated approach/framework to transform laboratory data in disparate US and EU databases into a common, standardized data structure. The results show promise in finding a way to use the data to address drug safety and other healthcare analytic questions. Due to the multiple components and overall complexity of laboratory data, researchers must be cautious and make the effort to understand the strengths and weaknesses of each data source that is contributing laboratory data to their analyses. Review by clinical experts in each of the countries represented is a critical component to a successful standardization process of laboratory data.

REFERENCES

FDA Sentinel Initiative. 2015. Sentinel Common Data Model – Laboratory Result Table Documentation. Accessed on 9 April 2021, from https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_Common-Data-Model_Laboratory-Result-Table-Documentation_0.pdf

Montis K, and Peil T. Multiplicative inverse or reciprocal. Accessed on 9 April 2021, from http://web.mnstate.edu/peil/MDEV102/U3/S23/S234.html

ACKNOWLEDGMENTS

The authors would like to acknowledge the following contributors to this body of work for their support: Kelly Castanos, Janine Collins, Robert Sharrar.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Irene Cosmatos, MSc
United BioSource LLC
Irene.Cosmatos@ubc.com