Status of clinical gene sequencing data reporting and associated risks for information loss

Douglas R. Mitchell *,1, Joyce A. Mitchell 2

Department of Health Management and Informatics, University of Missouri, 324 Clark Hall, Columbia, MO 65211, USA

Received 12 December 2005
Available online 15 March 2006

Abstract

Clinical gene sequencing is growing in importance and cost-effectiveness. In the past two years, the number of genes associated with disease has grown by roughly 25%. Knowledge of genetic variations will soon guide drug selection and dosages, predict risks from toxin exposures, and inform nutritional needs. Despite the significance of sequencing, methods for reporting results are problematic. Frequent use of paper and infrequent use of naming standards impede data exchange and make incorporation into the electronic medical record difficult. Reports often describe only variations found, rather than all data (all patient bases sequenced). Also, reports frequently do not describe reference data used to define variations. These practices create risks for loss of both data and information. Standardized electronic reporting of all data (all bases sequenced and all reference data) and electronic record systems capable of storing these results would both prevent data loss and simplify the preservation of information those data provide.

© 2006 Elsevier Inc. All rights reserved.

Keywords: Algorithms; Automatic data processing; Computational biology; Diagnosis; Computer-assisted; DNA/Analysis; Diagnostic use; Genetics; Genome; Human; Information storage and retrieval; Variation (genetics)

1. Introduction and overview

Approximately 7–9% of human gene loci can harbor variations known to have roles in disease [1]. As of March 2005, the Human Gene Mutation Database at the Institute of Medical Genetics indicates that 45,409 variations in 1766 distinct genes are associated with disease, and more are being discovered [2–4]. A similar number is obtained from a search of the National Center for Biotechnology Information (NCBI) LocusLink Database; as of November 6, 2004, there are 1703 known human genes associated with a disease [5–8]. In the past two years alone, the number of genes associated with disease has grown by roughly 25%, and more associations are being discovered at a rapid pace [9]. According to GeneTests, an online resource for medical genetics information, there are genetic tests offered for 1121 diseases as of April 8, 2005 [10,11].

Variations in the human genome not only are associated with disease but also affect health in other ways. Certain genetic variations have long been known to cause major defects in the human immune system that lead to devastating infections early and often in life; today, other genetic explanations are identified for more subtle differences in human susceptibility to infections by important agents such as the human immunodeficiency virus [12]. Variability in susceptibility to injury from environmental insults such as poisons and pollutants is largely hereditary, and evidence continues to mount that such risks may be knowable in the near future [13]. Such information could be a powerful tool of persuasion in preventing disease through behavior modification. For example, carriers of the Z allele of the SERPINA1 gene that causes alpha-1-antitrypsin...
deficiency have an increased risk for chronic obstructive pulmonary disease, and would be advised not to smoke cigarettes [14–17]. Similarly, knowledge of a person’s nutritional needs and an understanding of the likely health effects of his or her diet may soon be guided by genetic information [18]. Also, studies of the effects of genetic variation on drug mechanisms and metabolism will make the effects of prescription drugs on an individual more predictable. It will be easier to determine who is likely to suffer adverse effects from a drug, so there will be less abandonment of potentially useful drug development due to side effects rare in the general population [19]. Malignant hyperthermia (MH) is a life-threatening condition that occurs rarely after exposure to certain anesthetics. A person’s risk for MH can be assessed reliably but painfully, as it requires an invasive muscle biopsy. Attempts to develop a genetic test, however, are meeting with some success [20].

One method for identifying important genetic variations is clinical gene sequencing (more simply, sequencing). Sequencing identifies the bases of at least all coding regions of a gene for clinical purposes. For commonly used methods such as mutation scanning and analysis, the cost of testing a gene for all known variations rises with each new variation discovered. The costs of sequencing, in contrast have been falling [21], and methods are becoming more efficient. For example, L. Hood and others recently demonstrated sequencing of complementary DNA (cDNA) libraries at a rate of a million sequences per run [22,23]. Another advantage of sequencing over traditional methods for variation testing is that it characterizes bases at many more positions in the tested gene, the health implications of which might be someday realized without the need for additional testing. Sequencing will also aid in the recognition of patterns of rare variants associated with the growing number of diseases known to have complex inheritance [24]. Consequently, sequencing may soon be the most cost-effective means for most genetic testing. Already sequencing is gaining favor as a genetic test for disease: Of the human diseases for which genetic testing of any method were then offered, the availability of sequencing rose from 36% (373/1029) to 40% (435/1076) between August and November of 2004, according to GeneTests [10,11].

The first step in analyzing sequencing data is to classify each identified base as a variation or an expected base. The data used to make these distinctions are the reference data. The term reference data as used here is not synonymous with reference sequence. The latter refers to a unique string of characters representing bases, and possibly numbers indicating the nucleotide positions of those bases in their gene or chromosome. The best known human reference sequence database is the set of hand curated data maintained by the NCBI in the RefSeq databases [6,8]. The term reference data here more generally refers to all data used by the clinical laboratory as a basis for deeming a patient base a variation. Due to the presence in the human genome of many single nucleotide polymorphisms (SNPs) with no apparent phenotypic associations, often several bases may be considered expected at a given nucleotide position. Reference sequences as currently formatted in the RefSeq databases cannot convey a range of expected polymorphisms, and are therefore often only a portion of what comprise the clinical laboratory reference data. The RefSeq also do not contain the intrinsic sequences of genes; these data would be contained in the references data for a clinical laboratory.

The American College of Medical Genetics (ACMG) has recommended a minimum information set for sequencing reports that includes: nucleotide positions of variations, the fractions of exon and intron–exon boundaries sequenced, and interpretations of variations. Interpretations should characterize the likelihoods that the variations found cause the disorder in question, from “recognized cause” to “recognized neutral variant,” which in effect means that the ACMG recommends reporting all variations [25].

2. Current situation for reporting of clinical gene sequencing results

There are important aspects of sequencing results that the ACMG omitted from its recommendations and that in practice are often not reported. Rarely (and perhaps never) are all of the bases sequenced reported; in other words, the data generated are rarely entirely reported. The reference data are also rarely fully reported. A unique identifier for an entry in a reference sequence database (often NCBI’s RefSeq) or a PubMed identifier for published sequence is often, but not always, reported [26].

Given how results are often reported, questions arise. Could current sequencing reporting methods cause data and information loss? The genome is not fully understood, and future clinicians armed with a new understanding of the genome may need to know bases at positions currently being sequenced but not mentioned in reports. This paper examines whether it would be possible to reconstruct—to “reverse engineer”—all bases sequenced given the data most frequently being recorded, because if not, many gene sequencing tests may someday require otherwise unnecessary repeating.

Answering these questions first requires considering how to maintain sequencing data. The simplest way would be to put all data (every base sequenced) into the (preferably electronic) record. Alternatively, all bases sequenced in the patient’s gene could be deduced provided the record contains all variations and all of the reference data.

Data maintenance does not, however, ensure information maintenance. As they do for other tests, reference data for sequencing analyses change. For example, errors in reference sequences are discovered, and more accurate reference data for particular ethnicities are formulated. When the reference data change, the interpretations of patient data may need correction, as reported “variations” may in fact be expected and formerly unrecognized variations might exist.
If a complete record of all bases sequenced is lacking, then correcting the interpretation in light of changes to the reference data requires the following data:

All variations in patient sequence from old reference data version
AND [(old reference data OR new reference data) AND differences between old, new versions] OR (the old reference data and new reference data)]

Given that a sequencing report usually lacks complete reference data, even if the report contains all variations in the patient gene, not all patient bases sequenced can be deduced (resulting in data loss). Even if laboratories were fully reporting the reference data, the rarity with which they report changes to reference data means that necessary changes to the interpretations of patients’ results cannot be made. In other words, a risk for data (and, of course, information) loss is being created by the failure to report either (a) all of the bases sequenced; or (b) all variations and the reference data used. The risk for data loss would be averted if either (a) or (b) above were always reported, but a risk for information loss would remain because procedures to account for changes to reference data are absent. The answer, then, to the question “Could current sequencing reporting methods cause data and information loss?” is sadly “yes.”

Reference data used by laboratories are often based on reference sequence data from RefSeq at the NCBI [6,8], particularly cDNA reference sequences, current and all obsolete versions of which are maintained. As mentioned earlier, laboratories sequence more than just coding regions, and so the full laboratory reports would depend on reference data that are also applicable to introns and flanking regions—here termed genomic reference data (though they do not necessarily address every base in the gene). The use of genomic reference sequences can prevent problems that may arise with cDNA reference sequences due to multiple promoters and translation initiation sites, alternative splicing and varying poly(A) addition sites [27]. There are a number of public online databases from which genomic reference sequences can be downloaded using a browser, including the Nucleotide Sequence Database at the European Molecular Biology Laboratory [28], and the Genome Browser at the University of California Santa Cruz [29] as well as numerous locus-specific sequence databases. The RefSeq system does not store this genomic reference data.

It is possible, if not simple, to write scripts to automate downloading of some current genomic reference sequences from public databases. However, there are several impediments to the universal use of genomic reference sequences. Particularly when a gene is very large, the number of intronic (length) variants in the population may prevent specification of a single genomic reference sequence. The transcriptional start site may be debatable because the gene has several transcripts using different promoters and/or terminal exons, or the start site may be unknown [27]. For the purposes of our discussion, probably the most important point about genomic reference sequences is that there are no apparent means for obtaining obsolete versions from the databases mentioned.

3. Methods

To illustrate the difficulties presented by the current states of both sequence reporting methods and reference databases, we will here describe an algorithm that when given a “report” for one allele of a patient gene and a description of changes to the reference data since the report was generated: (a) derives all bases sequenced; and (b) makes any necessary changes to the variations identified in the report. Note, however, that several assumptions are made in devising the algorithm that are, in reality, rarely or never true (see Section 5).

The algorithm first accepts two “files” (Fig. 1). The first file (“the original report”) is partly comprised of the reference data, which consist of at least an identifier for a cDNA reference sequence from the RefSeq database. The reference data may also include what the laboratory considered expected differences from the reference sequence at the time of testing, which here will be termed additional reference data. The other data in the first file are: (a) what were considered variations in the patient gene at the time the results were first interpreted; and, importantly, (b) any patient data in addition to the variations needed to avoid ambiguity in reconstruction of all data (all bases sequenced), here termed additional patient data.

After accepting the two files, the algorithm reconstructs all data generated (all bases sequenced) at the time of testing. The remaining lines in File 1 indicate the original interpretation that two additional patient data

Original Report (File 1)

<table>
<thead>
<tr>
<th>NM_X.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,100</td>
</tr>
<tr>
<td>*28R</td>
</tr>
<tr>
<td>*32Y</td>
</tr>
<tr>
<td>59C-&gt;A</td>
</tr>
<tr>
<td>*62K</td>
</tr>
<tr>
<td>*62T</td>
</tr>
</tbody>
</table>

Lab’s Current Additional Reference Data (File 2)

| *28R |
| *32Y |
| *59M |

Fig. 1. Examples of inputted data. The first of the two inputted files contains the “original report” data. Note that all patient data describe one allele whereas real patient data would report two alleles. The first line in File 1 shows that the reference sequence version at the time of sequencing was NM_X.1. The second line indicates that positions 1 through 100 of the patient gene were sequenced. The lines beginning with an asterisk are the expected positions in the patient gene at the time the results were first interpreted; and, importantly, (b) any patient data in addition to the variations needed to avoid ambiguity in reconstruction of all data (all bases sequenced), here termed additional patient data.
testing. The first step in the reconstruction makes use of the website EFetch.

The website EFetch, produced by the National Center for Biotechnology Information (NCBI), allows for the online retrieval of cDNA reference sequence versions from the RefSeq database. It can be accessed using a browser, but has the distinct advantage of also allowing for automatically generated queries based on URL construction. By omitting the inputted identifier’s decimal place in its initial query to EFetch, the algorithm retrieves the current reference sequence version. For example, use of the URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=NM_000492&complexity=1&rettype=fasta retrieves from RefSeq the current version of the cDNA reference sequence for the Homo sapiens cystic fibrosis transmembrane conductance regulator (CFTR) in FASTA file format [30].

The first line of the returned FASTA file contains the identifier for the reference data version. The algorithm is able to extract this version identifier and compare it to the inputted identifier for the version used at testing. If they match, the algorithm notes that the reference sequence has not changed. If the identifiers do not match (as is the case in our example), the algorithm does a second query of EFetch, this time with the inputted identifier’s decimal place intact, thus retrieving the reference sequence used at the time of testing.

The algorithm next reconstructs all bases sequenced using the retrieved reference sequence used at the time of testing, the variations, and the additional patient data (Fig. 2).

Next the algorithm determines if changes to the variation data are necessary. The EFetch database is again queried, this time omitting the decimal place in the identifier, which results in the retrieval of the current reference data version, which is then compared to the old version.

The cDNA reference sequence versions may differ substantially; long stretches of nucleotides may be added to, or removed from the obsolete version. Unfortunately, RefSeq provides no clues as to the alignment of the versions; it numbers the nucleotides of each version irrespective of the other. What is needed, then, is alignment of the reference data versions. Fig. 3 shows part of the output of the program ClustalW in globally aligning the current and obsolete versions of a hypothetical cDNA reference sequence [30–33].

Alignment of the reference sequence versions allows for the stepwise checking for base discrepancies at each nucleotide position between the versions, and for the renumbering of nucleotides as necessary. Note in Fig. 3 the presence of 32 nucleotides at the 5’ end of the new version (upstream of the start codon) not included in the obsolete version; this explains why, in Fig. 1, the position at which a cytosine or thymidine may be expected that is reported in the additional reference data changes from 32 in the old report to 64 in the new report, for example. Also note that there was an error in the obsolete version at position 41 (73 in the new version), where it is now known that a thymidine is expected.

Lastly the algorithm makes corrections to the patient report using differences found between the reference sequence versions. In our example, much of the information in the report has changed (Fig. 4). The reference sequence identifier has been changed to that of the current version. The line identifying the nucleotide positions in the patient gene that were sequenced has been changed to reflect the numbering of the current reference sequence version. The additional reference data, additional patient data, and what are now interpreted to be the variations have also been updated.

4. Results

The demonstration of the algorithm illustrates how reports that do not contain all bases sequenced may
require data in addition to the reference sequence and all variations to prevent information loss. In our example, when the reference data changed one of the originally reported variations was found to be an expected base, and two additional variations were discovered.

As shown in Fig. 3, if the bases in the patient gene at the underlined positions had not been identified in the original report using the new reference data, the presence of 32 nucleotides at the 5' end of the new reference sequence that were not part of the obsolete reference sequence.

Fig. 3. A simulation of the output from ClustalW. Shown is a simulation of the output of the program ClustalW in globally aligning the current and obsolete reference sequence versions for NM_X. The position with large, bold characters is where the versions differ. Note the presence of 32 nucleotides at the 5' end of the new reference sequence that were not part of the obsolete reference sequence.

Fig. 4. (A) The reconstructed portion of the patient gene sequenced is shown next to the current reference sequence version. In boldface in the patient sequence are the variations after reinterpretation of the original report using the new reference data. Again, the large characters indicate the positions at which the base in the obsolete reference sequence version differed from that of the patient gene; and the underlined bases were deemed expected bases when the sequencing was done. The arrows indicate the position where the patient base was considered expected in comparison to the obsolete reference, but is now considered a variation. (B) The report is corrected to reflect the current reference data. (R = A or G; M = A or C; Y = C or T). See Fig. 1 for the original (now obsolete) report.

<table>
<thead>
<tr>
<th>A</th>
<th>Patient</th>
<th>1</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>33 ATGCAGAGGTCGCCCTCTGAAAGGCCCTGGTCCTCMAACTTTTTTCTGGTCCAAACTTTTTTT</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>New Ref</td>
<td>33 ATGCAGAGGTCGCCCTCTGAAAGGCCCTGGTCCTCAACTTTTTTCTGGTCCAAACTTTTTTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient</td>
<td>83 CAGCTGGAACAATTATTAGGAAAGGATACACAGACAGGCCCTTGAAAT</td>
<td>132</td>
<td></td>
</tr>
<tr>
<td>New Ref</td>
<td>83 CAGCTGGAACAATTATTAGGAAAGGATACACAGACAGGCCCTTGAAAT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>Corrected Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_X.2</td>
<td>33,132</td>
</tr>
<tr>
<td>*60R</td>
<td></td>
</tr>
<tr>
<td>60A-&gt;T</td>
<td></td>
</tr>
<tr>
<td>*64Y</td>
<td></td>
</tr>
<tr>
<td>604C</td>
<td></td>
</tr>
<tr>
<td>735-&gt;A</td>
<td></td>
</tr>
<tr>
<td>*91M</td>
<td></td>
</tr>
<tr>
<td>691A</td>
<td></td>
</tr>
<tr>
<td>96G-&gt;T</td>
<td></td>
</tr>
</tbody>
</table>
report, the algorithm could only have determined that the patient gene contains either of two bases considered expected at the time of testing. For example, either a guanine or a thymidine was expected at position 62 (later, 94) when the test was first interpreted. The provision in the report that a thymidine was found at that position in the patient gene thus prevented ambiguity during the reconstruction of all bases sequenced that otherwise would have caused data loss.

When the reference data changed, it became even more important that the original report's additional patient data preserved a record of all patient bases. At position 73 in the new reference sequence version there is a thymidine; there was an adenine in the corresponding position (51) in the old version. At this position in the patient gene is an adenine, and so when the results were first interpreted, it was not considered to be a variation. Additionally, at position 94 (previously position 62) in the patient sequence there is a thymidine that, although it differs from the guanine at that position in both the new and obsolete reference sequence versions, was not originally interpreted to be a variation. The laboratory then considered thymidine to also be an expected base. The laboratory now considers only a guanine expected at position 94, so the patient base is a new variation in the corrected report.

Also note that this example only uses the RefSeq data with cDNA values and does not delve into the complexities of other genomic variations including intron, promoter and flanking sequences. We omit this for two reasons. First, we omit this aspect because there are no publicly available records of changes in such reference sequences. Second, we omit this because the complexities of the current situation are already illustrated.

5. Discussion

Despite the importance of sequencing as a means for identifying clinically important genetic variations, common methods for reporting sequencing results present problems for the maintenance of these data and their associated information. While illustrative of these difficulties, the algorithm described above is not a working solution to these problems because it is founded on a number of assumptions that are often to almost always false in real clinical scenarios involving sequencing:

• That only exonic gene regions are of clinical interest,
• That only one allele is sequenced,
• That sequencing reports indicate the gene tested using a unique identifier for an entry in a reference sequence database,
• That variations found by clinical gene sequencing are stored and/or reported in a standardized, electronic format,
• That expected variations (known polymorphisms) are available in a standardized, electronic format,
• That clinical laboratories that use sequencing reference data sets generated in-house maintain records of them, including changes, in a standardized, electronic format.

The algorithm should therefore be considered a thought-experiment, and a call to action, but not a viable solution. We have chosen to illustrate it only in pseudocode, lest the availability of working code mislead the reader as to its purpose.

The algorithm does not, and cannot, retrieve genomic reference sequences, and this shortcoming alone makes the algorithm fail to solve the difficulties in reporting described. Instead, the reference data used in the algorithm (RefSeq cDNA reference sequences) are applicable only to gene coding regions. The reason for this, as discussed, is that there are no apparent sources of obsolete genomic references. In reality, laboratories most often sequence portions of flanking and intronic regions as well as exonic, and this practice will likely increase as the importance of non-coding gene regions is increasingly understood. Publicly accessible databases of past and present genomic reference sequences are badly needed.

Another limitation of the algorithm is that neither the inputted nor the outputted data include all salient features of sequencing reports. Especially important is that the algorithm is not intended to characterize the likelihoods that variations found are associated with the phenotypic change(s) in question, perhaps the most important facet of the interpretation of sequencing data.

The algorithm does, however, demonstrate that several aspects of RefSeq make it a model reference sequence database. RefSeq contains both current and obsolete versions of reference sequences in FASTA format. Retrieval of the current version can be done by manipulation of the identifier. An improvement to RefSeq would be the numbering of nucleotides in a way that provides alignment of reference sequence versions.

Another aspect of the example algorithm argues for reporting of all sequencing data. It would be at least doubly difficult to determine what additional patient data should be reported to allow for the future reconstruction of all bases sequenced when findings are being reported for both alleles, rather than for just one allele as in our example. Also, in comparison to reference and patient data likely to be used and produced in reality, the example sequencing data were very simple: no more than base substitutions for variations; and additional reference data comprised only of positions where more than one base could be expected. As the quantity and complexity of additional reference data used in the interpretation increases, so too would the quantity and complexity of the additional patient data needed in reports. For example, when sequencing is used to assess the probable severity and risk to proband relatives of a disease associated with repeat expansion mutations (e.g., myotonic dystrophy, Huntington chorea, fragile X syndrome), many repeating elements of broadly varying lengths could be expected.
Reference sequences comprised of simple character strings (as is typically the case) would be more informative if they used sixteen possible characters (\(2^4\)), one for each possible combination of expected bases or the absence of a base. Such reference sequences would eliminate the need to include in reports the positions and bases that might be expected in addition to those provided by traditional reference sequences and cover the frequently encountered insertions and deletions. Alternatively, as a source for reference data regarding expected polymorphisms, laboratories could make use of the NCBI’s SNP database (dbSNP) [34].

Several other problems with current reporting methods should be noted. Terms used for variations—conventions for numbering and naming of variation types (insertion, deletion, etc.)—differ between labs. Use of standard, human, and machine-readable variation terms would be preferable. Antonarakis described such standards [35], but they are used less by clinical labs than by researchers. The Cerner Corporation’s Clinical Bioinformatics Ontologyâ€”(CBO) proposes standards for variation terms similar to those of Antonarakis, and also standards for metadata describing the test methods in detail [36]. The CBO was only recently released to the public, however, and its acceptance by the laboratory community remains to be seen.

A more mundane but important problem with current reporting methods is the near-total reliance on paper, because the detail needed for gene sequence reporting is not part of most electronic medical records and the conveyance of the actual results is usually done by paper from the clinical reference laboratory. Entry by hand into an electronic record of all reported variations (not to mention all bases sequenced) by even the most careful operator would result in an unacceptable number of errors, were it (foolishly) attempted.

6. Conclusions

Sequencing as an initial laboratory technique for detecting genetic variations is gaining on more traditional methods. However, sequencing results are commonly reported using methods that make the incorporation and maintenance of this information in an electronic medical record difficult and error-prone. Unfortunately, the question “Could current sequencing reporting methods cause data and information loss?” asked earlier in this paper must be answered with “Yes.”

Clinical gene sequencing reports containing variations but not all bases sequenced (which are common) should contain all variations. Such reports should also indicate or provide the sequencing reference data used. Failing to do these two things seriously jeopardizes data integrity. However, as shown in the algorithm example, providing all reference and patient data does not alone guarantee data integrity. If there are data used in the analysis in addition to a typical reference sequence, such as positions where more than one base may be expected, reporting of patient data in addition to the variations may be necessary to avoid data loss.

Changes to reference data must be accounted for to correct errors in sequencing data interpretations. Obsolete reference data versions, and/or a “trail” of changes to reference data must be maintained. Laboratories using sequencing reference data sets generated in-house should also provide timely updates about any changes to them. Failing to do these things jeopardizes information integrity.

The clearest solution to many of the information problems presented by clinical gene sequencing results would be for laboratories to report all data generated (all bases sequenced) and all reference data (or identify their sources), and for those data to be recorded in the electronic record. The demand for reporting of all sequencing data could only be met by standardized electronic exchange of data, which unfortunately is rare today. Also unfortunate is that few electronic medical record systems are currently capable of receiving such data and storing them in a systematic manner.

The way sequencing is reported now would be like a lab reporting the results of Mr. Smith’s workup of analytes A through Z as “Analyte A is 2, with a reference interval of 3–10; Analytes B–Z: Normal” with no provision of the other reference data (Table 1a).

The next day the lab discovers that the reference intervals for analytes A, G, and Z depend on the subject’s age and gender (Table 1b). Mr. Smith’s A yesterday was actually normal. Unfortunately, however, the paper report has been sent and no one is able to contact Mr. Smith’s clinician. Mr. Smith’s G and Z may have been abnormal, but the lab has no way of checking since the actual values were not recorded in the laboratory system.

The laboratory neither reported all data, nor was prepared to report changes to the interpretation of those data. This would not be acceptable for reports of other clinical laboratory tests. Why should it be acceptable for reports of clinical gene sequencing?

Table 1
Mr. Smith’s lab report

<table>
<thead>
<tr>
<th>Analyte</th>
<th>12-Nov-2004</th>
<th>Reference interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) The results of Mr. Smith’s workup of analytes A through Z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>3–10</td>
</tr>
<tr>
<td>B–Z</td>
<td>Normal</td>
<td></td>
</tr>
<tr>
<td>(b) The day after Mr. Smith’s workup, the lab changes its reference data for analytes A, G, and Z. Mr. Smith’s laboratory test A was actually normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference Intervals (Updated 13-Nov-2004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New (age and gender-specific)</td>
<td>Old</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1–7</td>
<td>3–10</td>
</tr>
<tr>
<td>G</td>
<td>12–18</td>
<td>10–14</td>
</tr>
<tr>
<td>Z</td>
<td>46–68</td>
<td>35–56</td>
</tr>
</tbody>
</table>

Asterisk indicates value out of reference interval.
References


[3] Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff, University of Wales College of Medicine, <http://archive.uwcm.ac.uk/umcm/mg/hgmd0.html>; 2004 [Retrieved March 31, 2005].


