Human Factors Evaluation:
IV Premix Drug Labels Designed To Assist in the Reduction of Drug Selection Errors

Reducing Medication Errors by Optimizing Label Design

Medication errors, defined as “any error occurring in the medication-use process,” have been discussed in the literature in recent years. Of paramount concern is patient safety, which has prompted research into medication errors and corresponding efforts to reduce these errors. The medication process can be broken down into prescribing, transcribing, drug selection, dispensing, administering, monitoring, and documenting. Medication errors can occur at any of...
these steps. Therefore, optimizing a drug’s name, labeling, and package design has the potential to reduce medication errors. Efforts to reduce the potential for medication errors in the drug selection process include an assessment of the container label design.

**Global Labeling Environment**

Worldwide regulatory authorities responded to these trends by issuing new guidelines for the creation of proprietary names and readability recommendations for labeling. In an effort to mitigate errors, regulatory authorities in the United States and Canada require pharmaceutical companies to submit proposed proprietary names for review and to provide an analysis of the likelihood for confusion due to look-alike and/or sound-alike names. The European Union issued new readability guidelines, thus supplying the industry with a standard format for labeling to ensure label comprehension and the appropriate use of medicines by consumers and health care professionals. Additionally, the Australian Therapeutics Goods Authority recently published a review proposing regulatory standards for brand names, labeling and packaging, and requested industry and consumer input.

**Current Labeling Considerations**

When specifically considering the relationship between label readability and drug selection errors in the clinical setting, limited guidance is available. Current FDA labeling regulations in the United States outline a general format for the design of product labels consistent with United States Pharmacopeia (USP) guidelines. FDA regulations provide instructions for the presentation of the product name, strength, dosage form, and quantity, as well as identify inappropriate design features that would make the label misleading (eg, improper order and prominence of proprietary name, inadequate font size).

In 2010, the FDA held a public workshop to develop guidance for naming, labeling, and packaging practices intended to reduce medication errors. Such guidelines that take into account risk mitigation strategies are needed to help reduce the number of medication errors that continue to be reported to the FDA’s adverse event reporting system (Med-Watch) and through the USP-Institute for Safe Medication Practices (ISMP) Medication Errors Reporting Program. The FDA is collaborating with the pharmaceutical industry and organizations like ISMP to prepare recommendations that take into account human factors in the development of label design.

The objective of the labeling usability study presented in this report was to evaluate the effects of proposed label changes on drug selection errors for 6 intravenous (IV) premix drugs with practicing health care professionals.

**Study Methodology**

**Automated and Validated Test for Recognition**

The Automated and Validated Test for Recognition (AVTR) was developed using computer-simulated tasks designed to focus on label readability and drug selection errors in the dispensing and administration of drugs to patients in the acute care setting. Efforts were made to achieve the fidelity of a real-world setting by providing plausible nursing and pharmacy scenarios where the task of selecting premix drugs would commonly occur. Time pressure and background noise were incorporated to simulate aspects of the pharmacy and nursing settings. Confirmation bias, or the reader seeing what is familiar or what they want to see, rather than what is actually there, was built into the study design to approximate real-world selection errors.

**Study Design**

**Drugs Studied**

AVTR included the labels for multiple strengths of 6 unique IV premix drugs encompassing 20 drug product presentations.

**Participants**

This usability study was designed to focus on drug label readability and drug selection errors in the dispensing and administration of premixed drugs to patients in the acute care setting. Therefore, user groups were identified from typical...
drug selection process flows and clinical areas in acute care facilities that interact with the premix drug products within the scope of this study. Four user groups were identified:

1. Certified pharmacy technicians (from in-patient hospital pharmacies)
2. Registered pharmacists (from in-patient hospital pharmacies)
3. Critical care registered nurses (RNs) from intensive care units, emergency departments, operating rooms/post-anesthesia care units
4. Medical-surgical RNs

A minimum of 60 clinicians—15 from each user group—was required to complete the study. The participants were recruited from a variety of acute care hospitals. Although acute care environments and procedures may vary, the user groups had common user interactions with product labeling, including identifying a hierarchy of information to ensure correct drug selection.

**Computer-Based Study Design**

The participants in this study evaluated current labels and modified labels of the drugs using computer-simulated tasks with clinical context. An independent moderator oriented each participant to the study, demonstrated practice tasks to the participant if needed, and conducted participant debriefings.

A computer-based AVTR system was developed to present the computer-simulated tasks in each study segment and to facilitate participant debriefings. The AVTR system is comprised of a 30-inch, wide-screen WQXGA monitor, supporting software, and an input device. The dimensions of the monitor ensured that the label images were displayed in their actual size, with up to 3 products displayed side by side. Participants entered their drug selection choices using buttons on the input devices aligned with each label image on the screen. The AVTR system collected both participant responses and response times for each task in the study. The system also included a printer to print task response summary reports at the conclusion of each segment for use during debrief activities. A study session was comprised of the following:

- Two study segments: 1 self-paced segment and 1 time-pressure segment (quantitative summary and statistical analysis)
- Participant debriefings: 2 segment debriefings and 1 final debriefing (qualitative summary with a tabulation of observed errors and participants’ debriefing responses)

The study procedure flow for all user groups is illustrated in Figure 1.

**Test Segments**

The 2 test segments were representative of clinical practice for selecting drugs for dispensing (pharmacy participants) and selecting drugs for administration (RNs). During each segment, participants were presented with a stimulus (pharmacy labels for pharmacists and pharmacy technicians, and physician’s orders for RNs) and 3 drug label images presented side by side. The participants were asked to select the label that matched the stimulus.

In the self-paced segment, the stimulus information (pharmacy label or physician’s order) was displayed throughout a task and participants advanced to selection steps at their own pace. Consistent with routine situations, label images were displayed for as long as desired to enable a selection. A small percentage (10%) of the tasks presented a stimulus for which no match in the labels was displayed. The purpose of this small number of “no-match” tasks was to build the expectation that there will be a label that matches the stimulus, thereby contributing to confirmation bias. Figure 2 demonstrates the screen transition in the self-paced segment.
In the time-pressure segment, the stimulus information was displayed again throughout a task and participants advanced to selection steps at their own pace. However, to simulate emergency clinical situations, the label images were displayed for a very brief time (seconds) to force “quick glancing” of the labels. After the quick glance time period, images were removed during the remainder of the selection period. In this test segment, 1 of the 3 labels displayed always matched the stimulus; therefore, the “no-match” option was not present for any of the tasks. If participants have a confirmation bias, only allowing a quick glance at a label may increase the likelihood of error. Figure 3 demonstrates the screen transition in the time-pressure segment.

During each segment, the system paused to allow participants to rest. Recorded background noise of actual hospital pharmacies and nursing units was added to simulate a real-world clinical environment.

A debriefing was conducted at the end of each segment. The moderators asked questions to obtain the participants’ perception of their experiences and the possible reasons for any errors they made.

In the final debriefing activity, participants were presented with 1 label from each of the 6 IV premix drugs involved in this study (N=6 drug labels). The 6 drug labels were presented one at a time, with the associated current and modified labels.
Figure 3. Screen transition in the time-pressure segment.
displayed side by side. The participants were then asked to select the label that is more readable and would most likely allow a novice clinician to make decisions quickly and with fewer errors, and the reason for the choices. Their responses were captured by the AVTR system.

### Statistical Analysis

#### Task Response Error Rate

The individual number of errors was tabulated for each label type and by study segment and then combined to determine the overall error rate. For each study segment, logistic regression analysis was performed using SAS®21 procedure GENMOD assuming a binomial model with a log-link function (task response = 0 if correct ["Yes"] or 1 if incorrect ["No"])) to estimate the error rates along with the odds ratios of modified label relative to the current label.

Two-sided 95% confidence limits were calculated on the odds ratio and the associated $P$ value comparing the odds ratio to the hypothesized value of 1.0. A confidence interval on the ratio capturing 1.0 indicates insufficient evidence to declare label group rates statistically different. Conversely, a confidence interval on the ratio that does not capture 1.0 indicates sufficient evidence to declare label group rates statistically different at the 95% confidence level. Likewise, a $P$ value of 0.05 or greater indicates insufficient evidence to declare label group rates statistically different. Conversely, a $P$ value less than 0.05 indicates sufficient evidence to declare label group rates statistically different.

#### Task Response Adjusted Time to Correct Response

In order to penalize for an incorrect selection, the time to misidentification was censored with the maximum time to correctly identify a task plus a small constant of 0.1 seconds. Therefore, the misclassified time will always be higher than the time to correctly identify a task.

The distribution of adjusted time to correct response was positively skewed (ie, longer tail to the right). Taking the natural logarithm of the individual time to correct response resulted in a better approximation to a normal distribution (ie, time to correct response follows a lognormal distribution). The appropriate measure of central tendency of the lognormal distribution is the geometric mean.

For each study segment, a 2-sample $t$ test was performed to compare the label type means of adjusted time with correct response using the SAS® procedure $t$ test. The 2-sample $t$ test was performed on the natural logarithm (ln) of adjusted time to correct response. Comparing the means on the logarithmic scale is equivalent to comparing the geometric means.

The geometric mean for each task by label type, along with the mean ratios of modified label to current label was calculated. The modified and current geometric means were obtained by exponentiating the ln(modified mean) and ln(current mean), respectively. The ratio of the geometric means (modified label to the current label) was obtained by exponentiating the mean difference: ln(modified mean) – ln(current mean).

### Table 1. Participant Count Per User Group and Participant Characteristics

<table>
<thead>
<tr>
<th>User Groups</th>
<th>Certified Pharmacy Technician (N=16)</th>
<th>Registered Pharmacist (N=17)</th>
<th>Critical Care RN (ICU, ED, OR/PACU) (N=17)</th>
<th>Medical-Surgical RN (N=15)</th>
<th>Total (N=65)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age ≤45 years</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>32</td>
</tr>
<tr>
<td>Age &gt;45 years</td>
<td>6</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>Years experience (range)</td>
<td>2-23</td>
<td>2-40</td>
<td>6-40</td>
<td>3-39</td>
<td>2-40</td>
</tr>
<tr>
<td>English (native language)</td>
<td>9</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>45</td>
</tr>
</tbody>
</table>

ED, emergency department; ICU, intensive care unit; OR, operating room; RN, registered nurse; PACU, postanesthesia care unit
Two-sided 95% confidence limits were calculated on the mean ratio and the associated \( P \) value comparing the mean ratio with the hypothesized value of 1.0.

**Results**

Data from 65 clinicians, from 4 user groups, are summarized in this section. The prospectively defined number of participants for each of the 4 user groups was met. Table 1 displays a total count of participants in each user group and participant characteristics.

**Task Response Error Rate Results**

**Overall Error Rate**

The overall error rate was calculated by tabulating errors recorded by the AVTR system for each label type (modified, current) and by study segment. Table 2 provides the overall odds ratio comparison for the modified and current labels for the overall study segments (self-paced and time-pressure). Across both study segments, the overall error rate with the modified labels was significantly lower \((P<0.0001)\) than the overall error rate with the current labels.

Figure 4 illustrates the self-paced and time-pressure segments overall error rates for both the modified and current labels. The overall error rate for the modified labels was significantly lower than the overall error rate for the current labels for both the self-paced \((P<0.0001)\) and time-pressure \((P=0.0001)\) segments.

Additionally, the observed errors from the modified and current labels were assigned to a category of previously reported or not previously reported errors. In both study segments, the overall error rate for the modified labels was statistically lower \((P=0.0011 \text{ for the self-paced segment and } P=0.0005 \text{ for the time-pressure segment})\) than the overall error rate for the current labels for the previously reported errors. Furthermore, the modified labels did not introduce new errors that were not previously reported.

**Individual Error Rate Results**

When comparing the individual error rates for the modified and current labels of the 6 drugs tested, some variability in the results was observed. Across both study segments (self-paced and time-pressure), as well as for the previously reported or not previously reported errors, reductions in the individual error rates for some of the modified labels compared to the current labels were not statistically significant \((P>0.05)\).

**Task Response Time Results**

**Overall Time Results**

Table 3 provides the overall adjusted time to correct response comparison for the modified and current labels for both study segments (self-paced and time-pressure). Across both study segments, the overall geometric mean of adjusted time to correct response for the modified labels was significantly lower \((P<0.0001)\) than the overall geometric mean for the current labels with a geometric mean ratio of 0.87 (ie, the modified labels were processed 17% faster).

---

**Table 2. Overall Odds Ratio Comparison for Modified versus Current Labels for Both Study Segments**

<table>
<thead>
<tr>
<th>Study Segment</th>
<th>Odds Ratio(^a)</th>
<th>95% Confidence Interval on Odds Ratio</th>
<th>( P ) Value(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-paced</td>
<td>0.3901</td>
<td>0.3042-0.5003</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time-pressure</td>
<td>0.6763</td>
<td>0.5539-0.8259</td>
<td>0.0001</td>
</tr>
<tr>
<td>Overall</td>
<td>0.5392</td>
<td>0.462-0.6292</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

\(^a\) Odds Ratio = \((\text{modified label errors}/[\text{modified label tasks-modified label errors}] )/(\text{current label errors}/[\text{current label tasks-current label errors}])\). One may state with 95% confidence that the odds ratio is within the interval (confidence interval on odds ratio capturing 1.0 indicates insufficient evidence to declare label group rates statistically different).

\(^b\) Error rate = the number of errors divided by the total number of tasks. \( P \) value <0.05 indicates sufficient evidence to declare label group rates statistically different. An odds ratio less than 1.0 indicates the error rate for the modified labels is less than that for the current label.
label adjusted time to correct response was on average 0.87 of the corresponding current label time).

Figure 5 illustrates the comparison of the self-paced and time-pressure segments overall adjusted time to correct response for the modified and current labels. For both the self-paced and time-pressure segments, the overall geometric mean of adjusted time to correct response for the modified labels was significantly lower ($P<0.0001$) than the geometric mean for the current labels.

**Individual Time Results**

The individual geometric means of adjusted time to correct response for the modified labels for each of the 6 drugs tested was significantly lower ($P<0.0017$) compared with the current labels across both study segments (self-paced and time-pressure).

**Observed Errors and Debriefing Response Results**

As discussed previously, overall fewer errors were observed for the modified labels than for the current labels. In an evaluation of observed errors for each drug name, comparing the intended (correct) response against the actual (incorrect) selection, errors occurred more frequently among different paired strengths of each drug name (within-drug combinations) and with the “no-match” selection (participant selected “no-match” when there was a correct match shown). For all drug names, these errors occurred more frequently with the current labels and in the time-pressure segment.

When comparing within-drug combinations where errors occurred with the modified labels, participant debriefing responses were captured to determine the cause of the errors. Common explanations for these errors among all the drug names included the participant concentrating on the numerical values on the label instead of the units of measure, confusing mg and mcg, and confusion with identical layouts between strengths. These errors were more common in the time-pressure segment, which also included errors related to the participants not having enough time to make a selection and accidently selecting the wrong button.

**Root-Cause Analysis**

Over the 2-segment debriefings, participants were asked questions about the unique drug label combinations for the modified labels. If an error was observed, the participant was asked to discuss the reasons for the error. The information collected during these segment debriefings was used to analyze errors and identify the root cause of the errors. If no error was observed, the participant was asked if anything about the label
could potentially lead a novice clinician to make a selection error. The potential opportunities for error suggested by the participants focused on the font (size too small or not bold enough). Labels having one font color (particularly red) were described as difficult to read. Other potential opportunities included confusion with identical layouts between drug strengths, difficulty distinguishing concentration, and problems associated with proximity of the total drug amount and total volume.

**Final Debriefing**

In the final debriefing activity, participants were presented with each of the drug presentations, one at a time, with the associated current and modified labels displayed side by side. The participants were then asked to “select the label that is more readable and would most likely allow a novice clinician to make decisions quickly and with fewer errors.” The intent of this question was to capture the participants’ qualitative observations of label readability, and the response was captured by the AVTR system. Overall, the modified labels were selected 86% of the time for all drug names.

A final debriefing discussion was conducted from this activity to obtain clarity on the reason for the participants’ choices. Positive and negative comments provided by the participants for both the modified and current labels were evaluated for the most frequently stated reasons for selecting the modified or current labels. Overall, the modified labels had considerably fewer and less frequent negative comments reported.

**Discussion**

This labeling usability study demonstrated that overall, the modified labels resulted in fewer medication errors with reduced task response times compared with current labels. However, evaluation of the modified labels individually was observed to have varied results. Some of the modified labels did not have a significant reduction in drug selection errors in both study segments but did have significant reductions in task response times compared with current labels. None of the modified labels resulted in a statistically significant increase in error rate or task response time.

The study involved 65 practicing health care professionals, across 4 user groups, from acute care settings typically involved in the drug selection process. The participants covered a range of demographics: experience with IV premix drugs involved in this study, years of total clinical experience, native language, and age. Participants represented the user population of the tested premix drugs in the market.

The study method leveraged a computer-based system to present drug selection tasks representing routine and emergency clinical situations. Tasks were presented in a balanced and randomized fashion for both modified and current labels. Drug labels used with the tasks were selected for multiple purposes, including assessment of study responses in terms of previously reported and not previously reported errors. The computer-based system enabled collection of

---

**Table 3. Overall Adjusted Time (sec)\(^a\) to Correct Response Comparison for Modified versus Current Labels for Both Study Segments**

<table>
<thead>
<tr>
<th>Study Segment</th>
<th>Ratio</th>
<th>95% Confidence Interval on Ratio</th>
<th>(t) Test (P) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-paced</td>
<td>0.86</td>
<td>0.84-0.88</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Time-pressure</td>
<td>0.89</td>
<td>0.88-0.91</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Overall</td>
<td>0.87</td>
<td>0.85-0.88</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

\(a\) For incorrect responses, time is adjusted by taking the sum of 0.1 and the maximum time for a correct response, for each combination of intended drug and study segment.

\(b\) Geometric mean of time to correct response = the exponential of the mean of the log of the time to correct response. Ratio of geometric means is an estimation of the ratio of the geometric means (modified label to the current label) and is obtained by exponentiating the difference: \(\ln(\text{modified mean}) – \ln(\text{current mean})\).

A ratio less than 1.0 indicates the modified label geometric mean time to correct response is less than the current label.
participant task responses, response times, and debrief commentaries.

The system used for testing incorporated fidelity in several ways. The label images for the modified and current labels were developed to mimic bag labels in the actual environment in size and resolution. The proximity of the 3 labels displayed in a combination was representative of possible conditions found when selecting a drug label. The stimuli were created to be representative of physician’s orders and pharmacy labels seen in the acute care setting. Background noise was present to mimic the distractions found in the clinical environment. Moderators provided clinical context prior to the completion of segment tasks.

The overall quantitative data assessment demonstrates that the modified labels had a positive effect on error rates. The results of the usability study indicate the overall geometric mean of adjusted time to correct response for the modified labels was significantly lower than the geometric mean for the current labels for both study segments combined. The overall geometric mean of adjusted time to correct response for the modified labels was significantly lower than the geometric mean for the current labels for both the self-paced segment and time-pressure segment. Geometric means of adjusted time to correct response for the individual modified labels were also significantly lower than the geometric means for the individual current labels.

For the self-paced segment only, the response rate of “no-match” was analyzed. The “no-match” trials were included to facilitate the assessment of confirmation bias within the study. An attempt was made to provide the opportunity for confirmation bias from the perspective of what is within the control of the manufacturer. Physical factors that can contribute to confirmation bias, such as storage configuration, remain outside the scope of label design. The overall error rates with respect to confirmation bias were low for both label types. There was insufficient evidence to support the concept that the readability characteristics of the modified labels are different from current labels.

Figure 5. Overall adjusted time to correct response comparison for modified versus current labels.
labels with respect to the potential for overcoming confirmation bias.

Debrief responses were collected and analyzed for root cause. The key findings were reviewed for potential unacceptable residual risk. The overall conclusion is that the modified label design elements, raised by participants as the reason for observed errors or as potential for errors by novice clinicians, do not pose risk beyond the current labels. There were no major trends observed suggesting the potential for error. Furthermore, when assessing clinician preference, the participants in this study selected the modified labels at a higher frequency than the current labels.

The results from this labeling usability study were favorable and demonstrated that overall, the modified labels had an effect on reducing the occurrence of drug selection medication errors. The modified labels for the individual drugs in this study had variable results. However, the quantitative data combined with the qualitative data and root-cause analysis for each drug label resulted in the modified drug label being the preferred label for all drugs in this study. The FDA has accepted Baxter’s analysis of the results for the individual drugs; and as a result, all modified labels were approved by the agency.

The study had some limitations. For example, with a computer-based simulation, the stress of an emergency situation needed to be simulated with background noise and the added pressure of a rapid display of drug labels during drug selection. However, we were unable to account for each participant’s unique time interval that they would typically use when reading a drug label in an emergency situation. Based on information-processing principles, a specific time interval was calculated for the rapid display of labels in the time-pressure segment. This time interval was then used for all participants in the study. It is possible that this calculated time interval may not have been representative of the time all participants would give to reading a drug label in an emergency situation. In addition, due to the size of the monitor used in this study, we were limited to displaying 3 IV premix drug labels on the screen at one time. This was done in order to ensure that the label images were displayed in their actual size. However, in a true clinical situation, it is possible that more than 3 different IV premix drugs would be located in close proximity, adding even more complexity to the drug selection process.

Conclusion

In Baxter’s effort for continuous improvement, label design elements were used to modify existing IV premix drug labels with the intent to optimize label readability. Baxter then validated the optimized design through a usability study using an interactive computer-based test system. Although individual performance of the modified labels varied, combined results demonstrated that participants selected the modified labels more often and at a quicker rate with overall fewer errors than with the current labels. Furthermore, the modified labels did not introduce new errors that were not previously reported.

References


