

**INFORMATION SOCIETIES TECHNOLOGY (IST) PROGRAMME**



**HOMEY**

**“Home Monitoring through an Intelligent Dialog System”**

**DELIVERABLE: D19 (*public*)**

**WORKPACKAGE: WP8**

**Validation**

**D19 – Validation Report**

**Author:** Martin Beveridge (Cancer Research UK)  
Toni Giorgino (Consorzio Bioingegneria e Informatica Medica)  
Daniele Falavigna (Istituto Trentino di Cultura)  
Roberto Gretter (Istituto Trentino di Cultura)

**Submission Date: 30/9/2004**

**Partners: Engineering Ingegneria Informatica, (I), Reitek (I), Consorzio Bioingegneria e Informatica Medica (I), Istituto Trentino di Cultura (I), Cancer Research United Kingdom (UK), Language & Computing (B)**

## **SUMMARY**

This document is part of the result of the research project HOMEY funded by the IST Programme within the 5<sup>th</sup> Framework Programme as project number IST-2001-32434.

One of the goals of the HOMEY project is to develop a technology to be used for deploying innovative tele-medicine services. These new services will be based on an Intelligent Dialog System (IDS), designed and developed to effectively manage an incremental dialog between a tele-medicine system and a patient, taking into account user needs, preferences and time course of her/his disease.

The purpose of work package 8 is to evaluate accuracy and efficiency of the Intelligent Dialog System via show cases related to cancer care and home-monitoring of patients with chronic diseases. In all showcases a prototype will be run in order to assess performance and to derive potential useful indicators for improvement and exploitation of the technology.

The purpose of this deliverable is to describe the data collection and analysis of results for showcases in order to allow evaluation of the dialogue system as per the validation protocol described in deliverable D14.

# CONTENTS

<b>1</b>	<b>Abstract</b> .....	<b>4</b>
1.1	PURPOSE OF THE HOMEY PROJECT .....	4
1.2	PURPOSE OF WORK PACKAGE 8.....	4
1.3	PURPOSE OF THIS DELIVERABLE .....	4
1.4	LIST OF ABBREVIATIONS .....	4
<b>2</b>	<b>Introduction</b> .....	<b>5</b>
<b>3</b>	<b>Validation of Cancer Showcase</b> .....	<b>6</b>
3.1	OVERVIEW.....	6
3.2	METHOD .....	6
3.2.1	<i>Subjects</i> .....	6
3.2.2	<i>Protocol</i> .....	6
3.2.3	<i>Variables</i> .....	6
3.3	RESULTS.....	10
3.3.1	<i>Dialogue Manager Competence</i> .....	10
3.3.2	<i>Speech Recogniser Performance</i> .....	13
3.3.3	<i>Dialogue Manager Performance</i> .....	15
3.4	DISCUSSION .....	19
<b>4</b>	<b>Validation of Chronic Disease Showcase</b> .....	<b>21</b>
4.1	INTRODUCTION .....	21
4.2	ARCHITECTURE .....	21
4.3	EVALUATION.....	22
4.4	INTERNAL TRIAL.....	22
4.5	DATA COLLECTION ARCHITECTURE.....	23
4.6	CLINICAL TRIAL DESIGN.....	24
4.6.1	<i>Demographics</i> .....	24
4.6.2	<i>Call collection</i> .....	24
4.6.3	<i>Call diagrams</i> .....	25
4.6.4	<i>Procedure</i> .....	26
4.6.5	<i>Results</i> .....	27
4.7	DIALOGUE REFINEMENTS.....	27
4.7.1	<i>Problems encountered</i> .....	28
4.7.2	<i>Corrections</i> .....	29
4.7.3	<i>Evaluation</i> .....	30
4.8	DATA COLLECTED .....	30
4.9	RESULTS.....	31
4.9.1	<i>Grammar Coverage</i> .....	31
4.9.2	<i>Speech Recogniser Performance</i> .....	32
<b>5</b>	<b>Validation of the Multimodal Browser</b> .....	<b>35</b>
<b>6</b>	<b>References</b> .....	<b>38</b>
	<b>Appendix A: Script for Cancer Showcase Evaluation</b> .....	<b>40</b>

# 1 Abstract

## 1.1 Purpose of the HOMEY project

The purpose of the HOMEY project is to carry out research and develop technology to be used for deploying innovative tele-medicine services. The new services will be based on an Intelligent Dialogue System (IDS), designed and developed to effectively manage an incremental dialogue between a tele-medicine system and a patient, taking into account user needs, preferences and the time course of her/his disease. Intelligent dialogue requires the representation of goals, intentions, and beliefs about the effectiveness of the interaction in terms of quality of health care management. The dialogue system will require dynamic adaptation in order to understand the patient's medical problems and the physician's goals, handle misunderstandings, and carry-out argumentation regarding therapy options. In order to support such adaptation, a representation of the medical domain knowledge, the evolution of the disease of a specific patient, and the history of user-system interactions need to be represented.

## 1.2 Purpose of Work Package 8

The purpose of this work package is to evaluate accuracy and efficiency of the Intelligent Dialog System via prototype applications (showcases), which will be run in order to assess performance and to derive potential useful indicators for improvement and exploitation of the technology. The cancer showcases will allow evaluation of the possible roles of the *PROforma* knowledge model and the domain ontology in the speech interpretation process. The chronic disease applications target the task of collecting data from patients over the phone and are intended to investigate the clinical effectiveness of telemedicine systems for patient home monitoring.

## 1.3 Purpose of this Deliverable

The purpose of this deliverable is to describe the validation studies undertaken for both the cancer and chronic disease demonstrators.

## 1.4 List of Abbreviations

<i>CRUK</i>	Cancer Research UK	Partner responsible for WP6 & WP8
<i>ITC</i>	Istituto Trentino di Cultura	Partner responsible for WP3
<i>CBIM</i>	Consorzio di Bioingegneria e Informatica Medica	Partner responsible for WP4

## 2 Introduction

This deliverable describes the validation studies undertaken to validate the dialogue system developed in the HOMEY project for the showcases related to cancer care and home-monitoring of chronic diseases.

In the case of the cancer showcase, the system that was evaluated was a research prototype, as described in Deliverable D11 (Beveridge and Milward, 2003a), which was intended to investigate the use of existing knowledge representation schemas for medicine as the basis for dialogue management. This dialogue showcase was therefore built upon a pre-existing clinical application and so the evaluation that was carried-out was not one of clinical efficacy, as explained in D14, but rather to determine the practicality and usefulness of a speech interface to the underlying application. The system was therefore evaluated using measures such as speech recognition performance (i.e. word and sentence recognition accuracy), semantic accuracy, dialogue competence (range of dialogue phenomena handled), and dialogue performance as described in Section 3 of Deliverable D14 (Beveridge and Giorgino, 2003). The validation results for this showcase are given in Section 3.

In the case of the chronic disease showcase, the system that was evaluated was a mature dialogue-based clinical application, as described in Deliverable D6 (Stefanelli et al., 2002), which had previously undergone internal evaluation and refinement (see Deliverable D7; Stefanelli et al., 2003). For this showcase, the main objective of the validation study was therefore to determine the medical effectiveness of the developed systems in clinical trials. The validation results for this showcase are given in Section 4. This demonstrator was in Italian.

Both showcases have been validated using speech signals collected through the telephone platform described in the Deliverable D2 (Viganò, 2002). The cancer demonstrator was in English and the chronic disease demonstrator was in Italian.

The cancer showcase involved a limited number of English speaking people; nevertheless, as seen above, it has provided useful insights regarding both the feasibility and efficacy of a highly innovative dialogue management scheme based on the use of semantic knowledge.

The hypertension prototype has been used by a quite large number of real Italian patients (involved in the project through two Italian Hospitals: the Hospital of Pavia and the Hospital of Florence), who interacted with the system for providing both their clinical data and some information concerning their general health status. In this way, a large set of data has been collected that allowed to derive significant statistics with regard to speech recognition and dialogue performance, as well as to the clinical effectiveness of the system.

## **3 Validation of Cancer Showcase**

### **3.1 Overview**

In other work, CRUK has developed a system (ERA) for advising doctors on whether patients require urgent referral for suspected cancer (Bury et al., 2001). The system is currently accessed by a standard web browser that generates web pages for collecting patient data and reporting on results (see <http://www.infermed.com/era>). The cancer showcase used to evaluate the dialogue system developed as part of the HOMEY project uses the knowledge representation developed for ERA, along with an ontology provided by L&C, to provide a spoken dialogue interface for entering data into this system. For this showcase, the medical effectiveness of the underlying application has already been determined in previous studies (Bury et al., 2001) and so the evaluation here concentrates on the spoken dialogue interface.

### **3.2 Method**

#### **3.2.1 Subjects**

The validation study was based on dialogues by 6 users who ranged from experts on the task domain through to people with no specific knowledge of the domain or any wider knowledge of medicine or healthcare.

#### **3.2.2 Protocol**

Due to the technical nature of the domain for the Breast Cancer Referrals demonstrator, the validation study was based on scripted interactions (the script is given in Appendix A). Note, however, that users were asked not to just blindly follow the script but to ensure that the information acquired by the system was correct according to the described scenario. In fact, the script itself contained examples of correcting misunderstandings, using help etc. so users could quickly see how to use the system. The purpose of the script was therefore (a) to provide a scenario to be communicated and (b) to provide an example of how the system could be used, which the user could then generalise from in their own interaction. No specific training was provided.

#### **3.2.3 Variables**

The following variables were considered in deliverable D14 (Beveridge and Giorgino, 2003) as potentially important factors in evaluating the cancer referrals application: domain size & structure, degree of flexibility allowed at any point in the dialogue, verification strategy, variation in voices, and level of ambient noise. The following sections therefore describe the instantiations of these parameters in the evaluation study described here.

#### ***Domain Size and Structure***

The ERA domain consisted of three basic tasks:

- 1) acquire the data values required in order to make a referral decision
- 2) make a recommendation to the user, allowing them to query the arguments for and against different decision candidates, and confirm the final decision advised by the user
- 3) inform the user when the appropriate action (urgent referral, non-urgent referral, etc.) was complete.

The data acquisition task (task 1 above) required values for 16 data items to be acquired by the dialogue system. These are listed below.

Data Name	Data Type	Description
Age	Integer	The patient's age
Sex	Male/Female	The patient's sex
Nipple Discharge	Boolean	Whether or not the patient has nipple discharge
Bilateral Nipple Discharge	Boolean	Whether or not the patient has bilateral nipple discharge
Bloodstained Nipple Discharge	Boolean	Whether or not the patient has bloodstained nipple discharge
Cloth Staining Discharge	Boolean	Whether or not the patient has cloth-staining (i.e. large volume) nipple discharge
Breast Cyst	Boolean	Whether or not the patient has a breast cyst
Breast Lump	Boolean	Whether or not the patient has a breast lump
Breast Nodularity	Boolean	Whether or not the patient has a breast nodularity
Asymmetrical Nodularity	Boolean	Whether or not the patient has an asymmetrical nodularity
Intractable Pain	Boolean	Whether or not the patient has intractable pain
Acquired Nipple Deformity	Boolean	Whether or not the patient has an acquired nipple deformity
Gestational Nipple Retraction	Boolean	Whether or not the patient has a gestational nipple retraction
Nipple Eczema	Boolean	Whether or not the patient has nipple eczema
Breast Abscess	Boolean	Whether or not the patient has a breast abscess
Skin Ulcer	Boolean	Whether or not the patient has a skin ulcer

Note that there are ontological subsumption relations between some of the concepts associated with these data items:

- 1) *bilateral nipple discharge*, *bloodstained nipple discharge* and *cloth-staining nipple discharge* are all subsumed by the concept *nipple discharge*
- 2) *asymmetrical breast nodularity* IS-A *breast nodularity*
- 3) *breast nodularity* IS-A *breast lump*
- 4) *gestational nipple retraction* IS-A *acquired nipple deformity*

These ontological relations place constraints on the dialogue system as to the order in which to request items (e.g. it should ask whether there is any nipple discharge before asking whether there is any bilateral nipple discharge) as described in deliverable D11 (Beveridge and Milward, 2003). In addition, hypernyms and hyponyms of these concepts are included in the speech grammar for the user's utterances (see deliverable D9 (Ceusters et al., 2002)).

Once the above data has been collected, the system makes a recommendation to the user regarding the referral decision and allows the user to query the support for different candidates. These candidates and the arguments for each are shown below.

<b>Referral Candidate</b>	<b>Supporting Arguments</b>
Urgent Referral	Patient has acquired deformity of nipple Patient has breast nodularity Patient has nipple eczema Patient is thirty or older with a breast lump Patient has gestational nipple retraction Patient has skin ulcer
Non-Urgent Referral	Patient has cyst of breast Patient is fifty or older with nipple discharge Patient has intractable pain Patient has bilateral cloth-staining nipple discharge Patient has asymmetrical breast nodularity Patient is under thirty with a breast lump Patient has bloodstained nipple discharge Patient has abscess of breast
No Referral	None

If there are any arguments for urgent referral then that is recommended. If there are no arguments for urgent referral but there are some for non-urgent referral then that is recommended. If there are no arguments for either urgent or non-urgent referral then 'no referral' is recommended.

Once the user confirms one of the candidates the system carries-out the advised task (although, for the purposes of this demonstrator, no action is actually performed) and informs the user when it is complete.

The degree to which the domain size and structure affects the performance of the dialogue system largely depends on the degree of flexibility of user response that is allowed. The more flexible the range of possible responses, the more scope there is to make use of ontological relations in the dialogue planning.

### ***Degree of Flexibility of User Response***

In the current study, a wide form of mixed-initiative was used in which the user can not only answer more than one question with a single utterance, but can also explicitly take the initiative by asking questions of the system (e.g. clarification questions). The range of possible replies also includes references to over-specified and under-specified concepts (see deliverables D9 (Ceusters et al., 2002) and D14 (Beveridge and Giorgino, 2003)). This allows a complex dialogue structure as shown in the example below.

```
[1] S: What is the patient's age?
[2] U: she is 35, female, and she has a nipple disorder
[3] S: ok, what type of nipple disorder does the patient have?
[4] U: what do you mean?
[5] S: for example: nipple anomaly or nipple discharge
[6] U: she has some nipple discharge
[7] S: ok, ...
```

Here, when the user responds to the system's question regarding patient age in utterance [2], they also specify some additional information: the patient's sex and the fact that they have a nipple disorder. The concept 'nipple disorder' is under-specified and so the system responds in [3] by following-up this information with a clarification question. The user responds in [4] by issuing their own clarification request to the system and the system provides an appropriate response [5]. The user then answers the systems' clarification question [6], which is acknowledged by the system [7] before continuing with its dialogue plan.

### ***Verification Strategy***

Whilst, it would be expected that a high degree of flexibility in user responses would provide the most usable system, flexibility also requires a complex language model and so increases the chance of misunderstandings. Another important variable was therefore the verification strategy employed in order to handle mis-recognitions etc.

The verification strategy employed in this study was a 'batch' or 'global' strategy in which all the data required by the system in order to make a referral recommendation is collected without verification, and is then verified in a single batch before being used to make a decision. This is demonstrated in the example below.

```
[1] S: So this patient is thirty, female, and has a skin ulcer, intractable pain, a breast abscess, a breast cyst, and a bilateral nipple discharge. Is that correct?
[2] U: No, she isn't thirty
[3] S: What is the patient's age?
[4] U: Thirty-five
[5] S: Ok. So this patient is thirty-five, female, and has a skin ulcer, intractable pain, a breast abscess, a breast cyst, and a bilateral nipple discharge. Is that correct?
[6] U: Yes
[7] S: I recommend...
```

Here the system verifies all the collected data in [1], and the user indicates that part of the data is incorrect [2] leading the system to re-request that item [3]. Once the new value is acquired in [4] the system repeats the verification with the new data [5] which is confirmed by the user [6] and the system then uses that data to make a decision regarding its recommendation for referral or non-referral [7].

### ***Variation in Voices***

Speaker variation may also have an impact on system performance. In the current study all the speakers were male and only one had a non-RP accent. All spoke with standard prosody and at a normal speaking rate.

### ***Level of Ambient Noise***

The level of ambient noise in a particular environment is also a factor in system performance. In the current study the system was used in an office environment so there was a fairly high level of mainly unstructured noise (i.e. background noise such as doors opening and closing, typing, coughing etc) but also a small amount of structured noise (e.g. from other members of the office talking).

### 3.3 Results

The results of the validation study are broken into three main sections: an evaluation of *dialogue manager competence*, results for *speech recogniser performance* and results relating to *dialogue manager performance*.

#### 3.3.1 Dialogue Manager Competence

Two previous projects, TRINDI<sup>1</sup> and DISC<sup>2</sup>, have provided criteria for evaluating a dialogue manager's *competence* in handling certain dialogue phenomena. These are the TRINDI tick-list and the DISC dialogue management grid. Both of these are considered below.

##### *TRINDI Tick-List*

The TRINDI Tick-List (Bohlin et al., 1999) consists of three sets of questions that are intended to elicit explanations describing the extent of a system's competence.

##### PART 1: FLEXIBILITY OF DIALOGUE

The first set consists of eight questions relating to the flexibility of dialogue that system can handle.

1. Can the system deal with answers to questions that give more information than was requested?  
*Yes, the system can deal with extra information (mixed-initiative), and can also handle answers that are more specific than required (over-specified replies)*
2. Can the system deal with answers to questions that give different information than was requested?  
*Yes, the system can accept answers to any of the questions currently in the dialogue state, whether they were asked for or not. The supplied answers and any answers that are inferable from them (via ontological relations) are matched to the available questions.*
3. Can the system deal with answers to questions that give less information than was actually requested?  
*Yes, the system will raise one or more clarification questions in order to determine the more specific information that is required.*
4. Can the system deal with negatively specified information?  
*Yes. For example in the ERA domain the user can say that the patient "doesn't have a cyst" or that "she isn't forty" and so on.*
5. Can the system deal with 'help' sub-dialogues initiated by the user?  
*Yes, the system supports simple help sub-dialogues as follows:  
In response to a question by the system the user can ask for clarification (e.g. "S: does the patient have an acquired nipple deformity? U: what do you mean? S: for example: gestational inversion of nipple or nipple retraction").*

---

<sup>1</sup> Task Oriented Instructional Dialogue, European Telematics Applications Programme project LEA-8314.

<sup>2</sup> Esprit Long-Term Research Concerted Action No. 24823.

*In response to a proposal by the system the user can ask about the reasons for or against the system's recommendation and other presented candidates (e.g. "why do you recommend x?", "what are the arguments for y?", "are there any arguments against z?").*

6. Does the system deal with 'non-help' subdialogues initiated by the user?

*No, only help-type dialogues are currently supported.*

7. Can the system deal with inconsistent information?

*Data is always overridden by later information, so if the user, for example, specifies that the patient is age 30 and then later specifies that they are 40 then the earlier value of 30 is simply overridden (with no specific dialogue behaviour such as confirmation prompts). This is true even if the inconsistent values are in the same utterance, e.g. "the patient has a cyst and does not have a cyst" – in this case the system will record that the patient does not have a cyst.*

8. Can the system deal with belief revision?

*Previous information can be overridden with new information (as described above) but currently the consequences of any change are not calculated. E.g data that was inferred from the original information is not updated when the original information is changed.*

## PART 2: SYSTEM FUNCTIONALITY

The second set consists of five questions relating to the overall functionality of the dialogue system.

1. Can the system deal with noisy input?

*Yes, if the system is unable to match the input to the grammar then it informs the user that it was unable to understand their utterance and asks the user to repeat it.*

2. Can the system deal with barge-in input?

*Yes, the prompt playing is suspended and the grammar is activated for recognition.*

3. Can the system deal with no answer to a question at all?

*Yes, the system informs the user that it received no input and re-prompts them to answer the question*

4. Can the system check its understanding of the user's utterance?

*Yes, a verification strategy is used to ensure that all responses are checked before being used by the system. This strategy batches checks together so that all the information gathered concerning a particular entity is checked together (rather than checking understanding after each utterance). Responses are also checked if the information they specified is required by the system in order to proceed (i.e. it can go no further with only uncertain information).*

5. Does the system only ask appropriate follow-up questions?

*Yes, follow-up questions are only generated for information that is currently required according to the process specification, and for which answers have not been supplied and could not be inferred from other information (e.g. other answers).*

### PART 3: USE OF KNOWLEDGE

The third set contains just two questions relating to the ability of the dialogue system to make use of contextual/domain knowledge to provide appropriate responses to the user.

1. Is utterance interpretation sensitive to dialogue context?

*Yes, partial information is elaborated according to the last dialogue state. E.g. if the user says “she is 30” and the last question was regarding the patient’s mother’s age then the utterance is interpreted as meaning that the age of the patient’s mother is 30.*

2. Can the system deal with ambiguous designators?

*The system will assign an initial interpretation according to dialogue context, but it will be checked according to the verification strategy described above, so any misunderstanding due to ambiguity (amongst other things) can be corrected at that point. An underspecified reply may be ambiguous with regard to the specific data required by the system and leads to a clarification request immediately after the utterance, e.g. “U: my aunt had cancer S: by ‘aunt’ do you mean your mother’s sister or your father’s sister?”.*

### **DISC Dialogue Management Grid**

The DISC Dialogue Management grids (Heid et al., 1998) include a set of nine questions, similar to the Trindi tick-list above, that are intended to elicit some factual information regarding the potential of a dialogue system.

1. What initiative can the system cope with? (System/User/Mixed)

*The system supports mixed initiative.*

2. Free or bound order of main tasks?

*The state of the process specification at any point provides a partial ordering of main tasks. Further ordering is performed by the dialogue system according to ontological relations between the relevant task concepts, and other tasks (such as clarifications) may be added in response to current dialogue state.*

3. Does the system initiate repair dialogues?

*Yes, if no input, or input that doesn’t match the speech grammar, is received then the system informs the user and allows them to repeat their utterance. Before any information acquired by the system is used, it is checked according to a verification strategy (described above).*

4. Does the system initiate clarification dialogues?

*Yes, if an underspecified reply is provided by the user then the system will initiate a clarification sequence (as described above).*

5. Can the user initiate repair dialogues?

*The user can override previously provided information at any point by specifying new information, but they cannot initiate an explicit checking sequence, or query the information acquired by the system (e.g. “did I say the patient has a cyst?”).*

6. Can the user initiate clarification dialogues?

*The user can request clarification of a question asked by the system and the system then provides clarifying examples (e.g. “S: does the patient have an acquired nipple deformity? U: what do you mean? S: for example: gestational inversion of nipple or nipple retraction”).*

7. Can indirect speech acts be handled?

*There is no specific mechanism for interpreting speech acts (e.g. plan recognition). The conversational move associated with an utterance is specified as a semantic tag in the speech grammar, so, for example, “what do you mean?”, “I don’t understand” and “help” are all associated with a QUERY move. Hence the fact that the second two utterances are declarative and imperative respectively, and therefore only indirectly requests for information, is represented explicitly in the speech grammar.*

8. Is there any difference between the system’s use of speech acts and its ability to do topic spotting?

*Both speech acts and topics are represented as semantic tags in the speech grammar. E.g. the grammar rule that matches “she has a cyst” contains tags that assign something like the following structure: “(REPLY( she has a (CYST( cyst )CYST) )REPLY)”. Here “REPLY” is the move type and “CYST” is the recognised topic.*

9. Does the system deal with ellipsis?

*The system handles ‘short answers’, e.g. if the system’s question is “what is the patient’s age?” then any of the following are recognised replies “35”, “she is 35”, “the patient is 35”, “her age is 35”, “the patient’s age is 35”. The context of the question asked by the system is used to elaborate the partial replies, e.g. “35” is interpreted as “the patient’s age is 35”.*

### **3.3.2 Speech Recogniser Performance**

The following metrics were employed to evaluate speech recognition performance: word accuracy, sentence recognition, concept accuracy and semantic recognition.

#### ***Word Accuracy and Sentence Recognition***

Speech signals have been collected in two different periods: September 2003 and August 2004. In the intermediate period the dialogue management, based on the usage of domain ontologies, has been refined, mainly with the purpose of improving the semantic accuracy as will be explained in section 3.1.2. Speech recognition performance will be given for the two different data sets (i.e. **Sept2003** and **Aug2004**) both in terms of *Word Accuracy* (WA) and *Sentence Accuracy* (SA). WA is expressed (De Mori 1998) as:

$$WA = \left(1 - \frac{I + S + D}{T}\right) \times 100$$

where  $I$ ,  $S$  and  $D$  denote, respectively, the number of inserted, deleted and substituted words in the test set,  $T$  is the total number of words in the test set.

We point out that all of the speech signals composing the test sets, which means quite all of the recorded telephone signals, have been manually transcribed, in order to have the correct reference strings for the validation phase.

Table 1 gives the performance obtained on the *Sept2003* database, which consists of 285 telephone recordings, each corresponding to a sentence to recognize, for a total of 1011 words.

Table 1: *speech recognition performance obtained on the Sept2003 data set.*

<i>n. sentences</i>	<i>n. words</i>	<i>SA</i>	<i>WA</i>	<i>Errs ( D + I + S )</i>
285	1011	57.89 %	67.75 %	326 ( 48 + 98 + 180)

Note in Table 1 the larger number of substitution errors with respect to both insertions and deletions. This is probably due to “poor” coverage of speech recognition grammars, i.e. to the fact that the uttered sentences are not contained in the speech recognition grammars themselves. This has been demonstrated on a subset of sentences of the *Aug2004* database, as will be explained below. Note also the higher number of insertion errors with respect to deletions: this ratio could be reduced by properly tuning the language model probabilities assigned to the rejection networks (see Deliverable 5, Azzini et al., 2002, for the details) used inside the speech recognition grammars.

Experiments similar to those carried out on the *Sept2003* database have been performed on the *Aug2004* database. Resulting performance are given in Table 2.

Table 2: *speech recognition performance obtained on the Aug2004 data set.*

<i>n. sentences</i>	<i>n. words</i>	<i>SA</i>	<i>WA</i>	<i>Errs ( D + I + S )</i>
687	2459	59.24 %	71.82 %	693 ( 66 + 184 + 443)

Note that Table 2 exhibits performance trends similar to those of Table 1. In doing these experiments we noticed that a small set of speech files of the database (43 in total) are affected by spontaneous speech phenomena (e.g. coughs, laughers, etc), as well as by utterance truncation (e.g. “*yes and she has a []*”, “*she is thirty female and she []*”, etc) that slightly affect the overall performance. A deeper analysis of the speech recognition errors shows that most of insertion errors (184 in total, from Table 2) are due insertion of short words (this is typical of automatic speech recognizers) such as: *she* (22 in total), *he* (19), *is* (17), *ok* (17), *and* (13), etc. A similar problem has also been observed for deletion errors (66 in total from Table 2), where we measured 30 deletions of the word *a*, 6 deletions of the word *or*, etc. With regard to substitution errors, we noticed some substitutions between singular and plural words, namely *cyst* with *cysts* (12/26), *lump* with *lumps* (10/33) and *argument* with *arguments* (18/54). If we do not consider these substitutions as errors, WA increases from 71.82% to 73.49%.

Finally, as seen above, we measured the grammar coverage on a subset of speech files corresponding to 48 sentences among the 687 composing the *Aug2004* database. On this subset we measured a percentage of out of grammar sentences equal to 31.2%. This means that performance could be considerably improved by means of careful refinements of speech recognition grammars.

### ***Concept Accuracy and Semantic Recognition***

Concept accuracy measures the accuracy of the system in acquiring concepts (degree of semantic understanding) in a similar way to word accuracy, whilst the semantic recognition rate measures the percentage of completely correctly understood sentences (i.e. where every concept in the input utterance was correctly acquired).

For the *Sept2003* database, there were a total of 295 Semantic Units, of which 230 (77.97%) were correctly recognised, with 112 errors (insertions: 47, substitutions: 30 and deletions: 35) giving a concept accuracy of 62.03%. The total number of utterances was 285, of which 200 were correctly interpreted, giving a semantic recognition rate of 70.18%.

For the *Aug2004* database, there were a total of 1084 Semantic Units, of which 941 (86.81%) were correctly recognised, with 239 errors (insertions: 99, substitutions: 127 and deletions: 13) giving a concept accuracy of 77.95%. The total number of utterances was 687, of which 523 were correctly interpreted, giving a semantic recognition rate of 76.13%.

### **3.3.3 Dialogue Manager Performance**

The performance of the dialogue manager was evaluated using the following metrics: degree of success in achieving the desired task, cost of successful completion, and the overall usability of the system. These results are based on the data collected for the final system in August 2004.

#### ***Task Success***

The following aspects of task success were considered: the number of users who managed to complete a dialogue, the correctness of data acquired from user, and the correctness of data provided by the system (e.g. transaction success).

In the majority (80.77%) of cases, users successfully completed the dialogue. Hence, any errors that occurred were, in general, not severe enough to prevent the user from reaching the end of the dialogue. Of those cases where users hung up, about half were due to consistent mis-recognition of a single lexical item (“lump”) by the speech recogniser. The other half were due to mis-recognition of the user’s final decision as new data, causing the system to repeat the verification and decision stages and giving the impression that the system had started again from the beginning.

For those dialogues that successfully reached the decision stage (86.67%), the accuracy of the system in acquiring the data items necessary to make a decision (patient details, health symptoms etc as described above) was evaluated. Since a verification strategy was used by the system to check the correctness of these items, a high degree of accuracy was observed. In fact of the total of 416 data items for which values were acquired, only 10 values were incorrect, hence the overall accuracy was 97.6%. In the cases where incorrect values were acquired this seemed to be due to the user mis-hearing or not properly attending to the

verification prompt and confirming that data was correct even though there were errors. This may be due to the verification prompts being over-long or because of unclear pronunciation of some values by the speech synthesizer.

Finally, turning to transaction success, it was found that, of the dialogues that were completed by users, the referral task was successfully achieved (i.e. an appropriate referral was made for the patient being described) in 85.71% of cases. Unlike the acquisition of patient and health symptom data, the referral decision acquired from the user was not verified by the system and so, in the cases where an incorrect referral was made, the error was generally due to a mis-recognition by the system during the negotiation of the decision (e.g. interpreting a user utterance as confirming a candidate when it was actually a question regarding the arguments for that candidate). The success of the verification strategy for data acquisition, however, suggests that it should be extended to also include verification of decisions.

### *Dialogue Costs*

The task success of a dialogue system needs to be weighed against the costs in using the system. For example, a system that verified every data item immediately at the point where it was acquired would have a high task success but at the cost of a long and tedious dialogue. The following measures of dialogue ‘cost’ were considered: system response time, overall amount of time required to complete a dialogue, the number of turns required to complete a dialogue, and the proportion of turns that were spent correcting errors such as misrecognitions, misunderstandings etc.

The median response time for the dialogue manager (from receiving a voice browser request to sending a response) was 531ms (ranging from a minimum of 16ms to a maximum of 12047ms). High response times occurred only at the start of a dialogue in cases where the system had previously been reset and so the domain ontology and cache had been removed from memory. In these cases, this data had to be reloaded from file. Once loaded, however, the system response time returned to average levels, and subsequent dialogues were at this level throughout (the cache persisted until the server was reset).

The median total number of turns (including both user and system moves) was 50. This is higher than for some other reported systems (e.g. the DARPA Communicator Demonstrator (Walker et al., 2000)), but the task being performed here is also more complex than those described in most evaluation studies, which are typically based on simple information-lookup tasks (e.g. the DARPA system mentioned above is a travel planning system).

The median time taken to successfully complete a dialogue was 277s (i.e. 4 minutes and 37 seconds) and ranged between 188s and 444s. This metric correlates closely with the number of turns per dialogue (correlation coefficient  $r = 0.88$ ). The fastest dialogues were those where there were few mis-recognitions (and hence a lower number of corrections) and where the user took advantage of mixed-initiative to provide a lot of information in one utterance (e.g. “she is thirty, female and has a bilateral nipple discharge”) rather than waiting to be prompted for each item. Conversely, the longest dialogue was one in which the speech signals were very noisy and there were many speech recognition errors (and hence corrections).

This dialogue duration metric becomes more meaningful when normalised according to the complexity of the dialogue, e.g. as measured by the average number of concepts acquired. In this study, the median number of concepts acquired in an interaction was 37 (in a range

between a minimum of 13 and a maximum of 58, with the lowest values recorded for dialogues that were not successfully completed). Hence, the median time to acquire a concept can be estimated as 6.95s per concept (ranging between a minimum of 4.62s and a maximum of 11.50s). These times include both user and system turns, and so it appears that the data acquisition accuracy reported earlier was not achieved at the cost of dialogue efficiency as measured by dialogue duration.

Finally, turning to the correction rate metric, it was found that the median number of turns involving spontaneous (i.e. non-scripted) error corrections in each dialogue was 2, and that the median proportion of turns spent correcting errors was 8.20%. Hence, the dialogues also had a low cost in terms of the amount of time spent by the user in correcting system errors.

### *Usability*

The following usability metrics were considered: the number of times a user made use of 'help', the quality of system responses, the quality of user responses, and user report (e.g. would they use it again etc.)

In general, there was no spontaneous (i.e. non-scripted) use of system help during a dialogue (median number of help utterances was zero). In fact, help was only used 9 times in the entire set of dialogues. For those dialogues in which it *was* used, the median proportion of turns spent on help in a dialogue (a notional 'help rate') was 5.3% (ranging from a minimum of 2.56% to a maximum of 10.53%).

The quality of system responses was evaluated according to the SUNDIAL 'Contextual Appropriateness' metric (Simpson and Fraser, 1993). Each response was labelled as either appropriate (AP), borderline appropriate/inappropriate (AI), completely inappropriate (IA), incomprehensible (IC) or a total failure (TF) (e.g. failure to respond at all to the user's utterance). It was found that, of 691 system responses, 547 (79.16%) were classified as AP, 32 (4.63%) as AI, 100 (14.47%) as IA, 8 (1.16%) as IC, and 4 (0.58%) as TF.

Most of the responses in the IA category occurred as a result of mis-recognitions of the user's last reply (e.g. recognising "she has an ulcer" as "she has a lump" and then asking "what type of lump does she have?"). Of those responses included in the AI category, most were cases where the response was appropriate according to local context (e.g. the user's last utterance) but not really appropriate in the wider context (the current dialogue segment). This is illustrated by the following fragments (taken from the dialogues recorded in this study).

```
[1] S: I recommend urgent referral ... What do you advise?
[2] U: ok urgent referral
[3] S: So this patient is 2. Is that correct?           IA
[4] U: no she is forty
[5] S: So this patient is 14. Is that correct?        AI
[6] U: no
[7] S: In what way is it incorrect?                  AI
[8] U: she is forty
[9] S: So this patient is 40. Is that correct?       AI
[10] U: yes
```

In this fragment, the system is in the decision phase of the dialogue [1] and the user advises that an urgent referral is necessary [2]. Unfortunately, the user utterance is mis-recognised as a correction to the patient's age (in this case the string returned by the speech recogniser was

“ok her age is two or”) and the system then tries to verify the new data [3]. This system response is therefore inappropriate (IA) in the context of the user’s last utterance. However, the user then goes on to correct the misunderstood data [4, 6, 8, 10] which entails further system responses [5, 7, 9] which are appropriate in the local context given that the user now really is changing the age value, but are inappropriate given the wider context of confirming a referral decision (AI).

The following dialogue fragment provides a similar example.

[1] S: What kind of lump does the patient have?	<b>IA</b>
[2] U: no lump	
[3] S: And is it a breast nodularity?	<b>IA</b>
[4] U: wh...what do you mean	
[5] S: For example: pre-existing breast nodularity, or symmetrical breast nodularity	<b>AI</b>
[6] U: Neither	
[7] S: Does the patient have a skin ulcer?	<b>AP</b>

In this fragment the system has mis-interpreted a user utterance as meaning that the patient has a lump [1] and the user is trying to correct the data [2]. Unfortunately, the user utterance is mis-recognised (the string returned by the speech recogniser in this case was “a lump”) leading to another inappropriate (IA) utterance [3] trying to clarify the type of lump. The user then asks for help [4] and so the system response [5] providing an explanation is appropriate in the local context of the user’s last utterance but inappropriate in the wider context in which the user is trying to indicate that the patient does not have lump at all (AI).

It is possible to combine some of the categories given above in order to get a broader characterisation of system responses. For example ‘adequate responses’ could be defined as those that are contextually appropriate or borderline appropriate (AP + AI), and ‘inadequate responses’ could be defined as those that are inappropriate, incomprehensible or failures (IA + IC + TF). Given these categories, 83.79% of responses would be classified as adequate and only 16.21% as inadequate. Hence, the majority of system responses were sufficiently appropriate (given the context) to be understandable by the user (and hence most dialogues were successfully completed, as described earlier).

As well as evaluating system responses, the quality of user responses was also analysed using the Behavioural Coding Scheme (Sutton et al., 1995). This measures the degree to which user answers could be characterised as ‘concise and responsive’ (e.g. “S: what is the patient’s age? U: 35”), ‘usable but not concise’ (e.g. “S: what is the patient’s age? U: her age is 30”), ‘responsive but not usable’ (e.g. “S: what is the patient’s age? U: she’s middle-aged”), ‘not responsive’ (e.g. “S: what is the patient’s age? U: I don’t know”), or as containing no speech (e.g. just noise or silence).

It was found that, of a total of 687 user responses, 347 (50.51%) could be classified as concise and responsive (C), 300 (43.67%) as usable but not concise (U), 14 (2.04%) as responsive but not usable (R), 3 (0.44%) as not responsive (NR) and 23 (3.35%) as not containing any speech (NS). Note that a very high proportion of category C compared to category U would be characteristic of a primarily system-initiative dialogue with the user simply providing answers as briefly as possible. On the other hand, the pattern observed here of almost equal numbers of category C and U responses is indicative of a mixed-initiative system in which the user can provide more data than was requested, phrase it in more

complex ways than just a single-word reply, and ask clarificatory questions of the system. Hence, the analysis of user responses is indicative of a high degree of dialogue flexibility.

The first three of these categories (C + U + R) can be amalgamated into a single class of 'Adequate Answer' and the last two (NR + NS) into a class of 'Inadequate Answer' (Sutton et al., 1995). Under this scheme, 96.22% of user responses were adequate and only 3.78% were inadequate. Hence, the prompts provided by the system were sufficient to elicit adequate (responsive, usable and/or concise) answers from the user.

The final metric typically employed to measure usability is 'user report'. Since the cancer showcase was a prototype system, however, no systematic investigation of users' qualitative assessments of usability was undertaken. However, anecdotally, users did report finding the system reasonably easy to use and, in particular, were impressed with the ease with which they could correct incorrect data items, e.g. using a single utterance such as "no she is forty and she doesn't have a cyst but she does have an ulcer".

Almost all users also reported problems with repeated mis-recognitions of certain words, which required some perseverance to get past, and problems with the lack of verification of the final referral decision (as noted earlier). The last problem was compounded by the fact that all data slots were kept open (i.e. fillable) until the end of the dialogue, so last-minute errors could cause previously set slots to be overwritten and the verification and decision phase of the dialogue to be repeated just when the dialogue appeared to be complete. However, except for a small number of cases, these problems did not prevent the user from completing the dialogue.

### ***Reconfigurability***

The reconfigurability of the dialogue system was tested by porting it to a new, larger, and more complex domain, namely genetic risk assessment (RAGs). Despite the fact that the original system, as developed for the breast cancer referrals domain, took about 2 person years to develop, it took only 3 person months of effort to extend the system to support the RAGs domain. Although, the RAGs application has not yet undergone any performance evaluation, this alone suggests that the original framework does generalise to different and more complex medical domains, and that the infrastructure of the existing system is sufficient to allow new domains to be quickly implemented.

## **3.4 Discussion**

The results presented here for the cancer showcase, whilst preliminary in nature, suggest that we have developed a framework for rapid implementation of dialogue systems based on high-level knowledge representations, namely: a process specification (e.g. represented in the *PROforma* language (Fox et al., 2003)) and a concept ontology (e.g. the L&C ontology (Ceusters et al., 1999)). Certainly, we have been able to implement a new and larger application than the one evaluated here (and indeed than many of the systems reported in the dialogue literature) in relatively short amount of time. This was possible because much of the expertise that was gained in building the original application has been formalised and captured in the infrastructure of the generic dialogue system, allowing it to be subsequently re-used.

Furthermore, analysis of the competence of the breast cancer demonstrator indicates that applications implemented using this framework can handle a wide range of dialogue

phenomena, some of which are not supported even by state-of-the-art systems. For example, the use of ontological information to dynamically re-order tasks for maximal coherence is not supported by TRIPS, GoDiS or HMIHY (see (Milward and Beveridge, 2004) for a more detailed discussion). This high level of competence derives from basing the dialogue system on high-level knowledge representations, which allow more sophisticated reasoning about dialogue structure than the simple task lists typically employed (e.g. the agenda in GoDiS).

Finally, our evaluation of the overall dialogue system performance, though tentative, suggests that these benefits have not been gained at the expense of performance. In fact, the cancer application demonstrates good speech recognition performance (concept accuracy of 77.95% with 86.81% of concepts correctly understood), and good performance in acquiring data (97.6% correct) and successfully completing transactions (80.77% of dialogues completed, of which 85.71% achieved the dialogue goal). Moreover this was achieved whilst maintaining a fast response time from the dialogue manager (on average 531 ms from request to response), and an efficient dialogue from the user's point of view (on average 6.95s to acquire a concept, including both system and user turns) with a low correction rate (on average 8.2% of turns spent correcting errors).

In terms of usability, the majority of system responses were found to be contextually appropriate (79.16%, with a further 4.63% borderline cases), and elicited user responses that were almost all (96.22%) adequate (i.e. responsive, usable and/or concise). Possibly for these reasons, users did not, on average, make use of system help, and when they did it was only for a small proportion (5.3%) of turns. Although users' own assessments of usability were not formally investigated (e.g. via user questionnaires), anecdotal report suggested that users generally found the system easy to use, although it was also clear that there areas in which the system needed to be improved, in particular: verification of decisions, and more robust handling of repeated speech mis-recognitions of the same concept. These are both also borne-out by the quantitative performance data reported here, and are therefore topics for further work.

## 4 Validation of Chronic Disease Showcase

### 4.1 Introduction

The evaluation of the Chronic Disease demonstrator was conducted in two steps. First, an internal trial carried out with the help of volunteers has addressed the main technical and usability issues. The internal trial was, in its turn, divided into two branches. The results of the internal trial have been described in a purposed previous deliverable.

To assess issues involved in the introduction of the web-based clinical record, augmented by the automated adaptive dialog system, we arranged collaboration with major Italian institutions devoted to the care of hypertensive patients (Ospedale Careggi of Florence and Ospedale S. Matteo of Pavia). The physicians have been involved both in the design of the Electronic Health Record and the organization and refinement of dialogue structure and steps.

In August 2003, the preparations for the clinical trial were finalized, and the trial started. The study foresaw the split of patients into two groups (control and treatment). Patients were enrolled progressively by collaborating physicians, after being requested to sign the customary *informed consent* letters. The technology and clinical evaluation process for the Chronic Disease demonstrator therefore took place in a *real* setting involving actual patients, who had not been specifically trained for using the telephone system. Age distribution and further demographics are given later in this section.

Evaluation results collected from the clinical trial of the Chronic Disease demonstrator have been analyzed along the following major directions:

1. Verification of technical aspects of the Automatic Speech Recognition (ASR) technology chosen, grammars, and the rest of the interaction architecture
2. Assessment of effects of the combined system (web-accessible database plus home monitoring service) on patients' clinical state and perception of selves' health.

### 4.2 Architecture

The dialog system is interfaced to the electronic health record (EHR) of each monitored patient. Two actors are therefore allowed to enter data into the EHR. In the first place, the physician may use a conventional (web-based, keyboard and mouse) interface to store and update patient information. On the other hand, the patient is also allowed to enter the self-measured data by the means of a telephone. Figure 1 illustrates the data flow inside the application. Patients periodically call a dedicated telephone number and engage a dialogue with the system, which talks and interacts with them to acquire clinical data, monitor their style of life and ask about the occurrence of possible side effects.

Values entered by the patient are distinguished by those taken during encounters by a special icon that appears near them on the graphical interface. Data are entered in tables through a graphic user interface, whose design has kept into account the suggestions of the clinicians (figure 1). The database holds several values that will affect the next dialog session: e.g, whether the patient has been prescribed a diet, his current weight, the date of the next visit. When a new call is set up for a specific patient, values are extracted from the database and the corresponding call is prepared. The software performing this task is called "adaptation

agent”; part of the adaptations are in fact performed by the agent to try to make the dialog more effective and friendly.

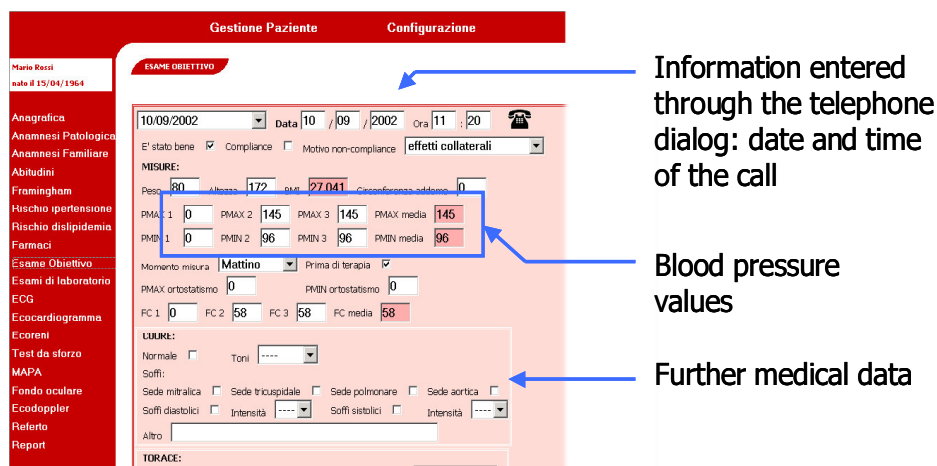
### 4.3 Evaluation

Despite the care put into the design of a dialogue application, questions and grammars, the fact that untrained users interact with a complex system poses remarkable usability challenges. The dialog application should be easy to use and understand, and robust. A number of issues in the design of our application were therefore analyzed by the means of on-field trials. The objectives of the evaluation phases were:

- *Testing the reliability of the system* – Recognition errors, although annoying, should not affect the user’s ability of proceeding in the remainder of the dialog.
- *Extension of grammars and lexicon* – Grammars should capture most of users’ answer schemes.
- *Reformulation of questions’ wording* – Users’ answers wording is influenced by how questions are asked.
- *Extraction of patient’s learning curve* – To address adaptability issues, one desires to put each patient into an ability class. It is then possible to study how quickly users learn to use the system, so they can be hinted about more advanced features, like mixed-initiative.
- *Assessing the clinical effectiveness* – The ultimate goal of the project is both to raise patient compliance with the guideline and to expand the availability of data for the use of the clinician. Collecting quantitative information involving real patients may assess whether the system helps to achieve a better quality of life.

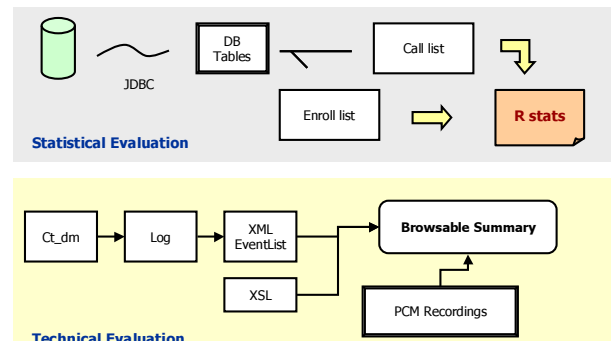
### 4.4 Internal trial

To debug our dialog application according to the goals listed, we designed an evaluation plan to happen in two phases. The first phase (*internal trial*) involved a group of volunteers, which were assigned a realistic disease profile. They were asked to call the system at fixed schedules, pretending to be hypertensive patients. Profile assigned included therapy, average blood pressure (BP) values and so on; the profile, in turn, affected the inquiry of side effects during the dialogue. The volunteers were asked to annotate and report problems and obstacles found. The call collection ended after every user performed the assigned number of calls and therefore addressed technical and usability issues. It involved approximately 15 people and collected 150 dialogues, amounting to about 500 minutes of conversation, of which 150 were human speech, the rest synthesized by the text-to-speech component (TTS).



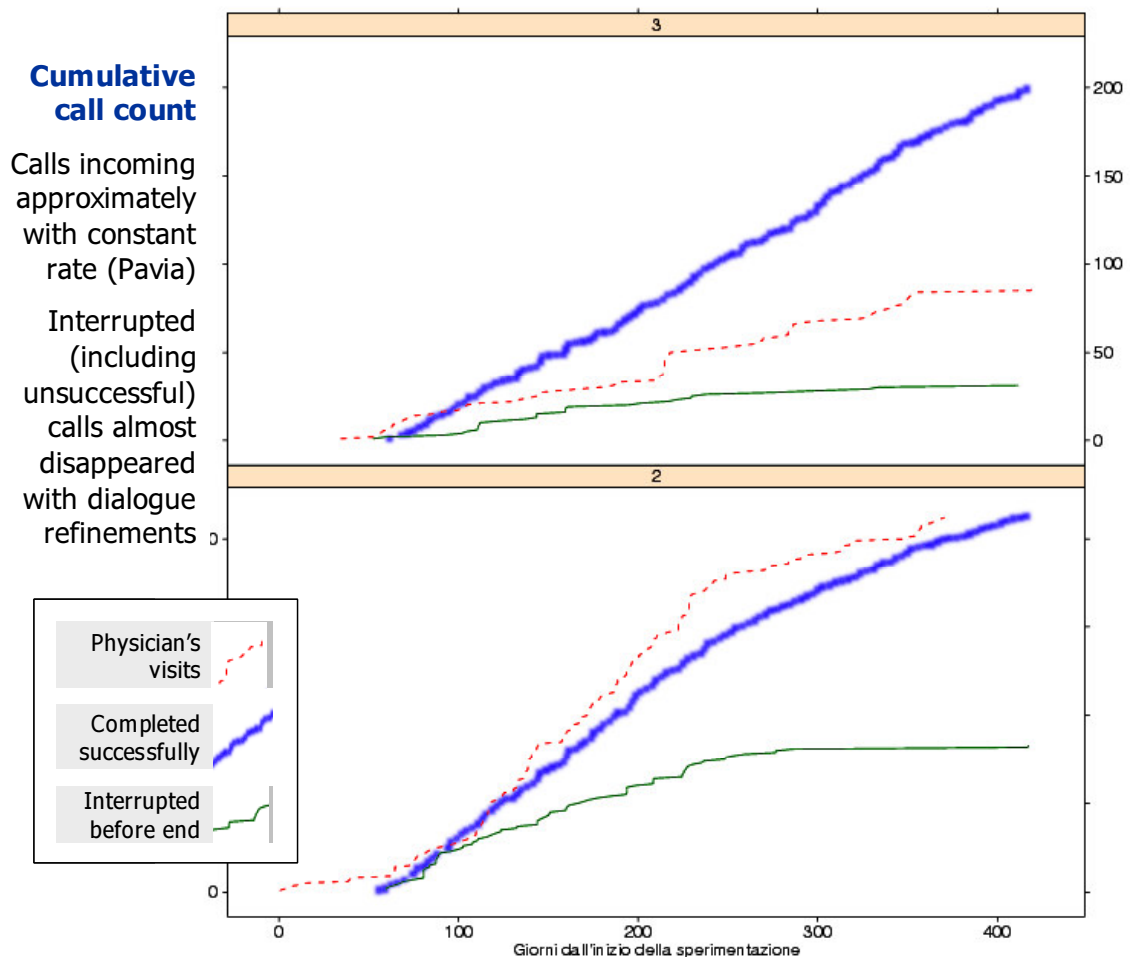
## 4.5 Data collection architecture

The procedure designed to collect and analyze data, coming from the real users' calls, keeps into account two sources of information: (a) call logs, and (b) clinical information about the patients. The former is gathered from information generated at run time by the computer telephony platform. By post-processing the logs, it is possible to automatically gather the detailed per-call measures listed in figure 3.



The other valuable source of information is the EHR itself, where a record is made every time a patient uses the voice system to enter his data, or doctors login via the web to store outcomes of real encounters. Contrarily to the telephony logs, information in the EHR has a coarser granularity: it only stores confirmed values, and not, for example, misrecognitions, repetitions and other phenomena, which are dealt with within the dialogue system.

Figure 3 shows a part of the numeric quantities that can be gathered by processing runtime logs. The availability of detailed records of such quantities is mostly useful for the assessment of system quality. To start, one could monitor the fraction of confirmation questions that received negative answers, discovering possible relation to changes made to the system or even demographic data. One exciting prospective use, related to this one, is to predict the behavior of the system with a specific user, and change dialogue strategies accordingly. Given the availability of call data, one could think of applying knowledge discovery techniques to this task. Such techniques would be most useful if an independent



rating of the dialogue quality is available, such as a measure of “successfulness” obtained subjectively by a human rater, which reviewed the recorded conversation. Further investigation on this technique is ongoing (8).

Start date / time	
Call duration	Number of prompt sessions
Whether user was anonymous	Time spent in recognition mode
Patient code	Time user actually spent speaking
List of concepts acquired	Time for pronouncing prompts
Number of fields acquired	Whether call reached bye message
how many were unique	How many times help was requested
Number of recognition sessions	How many rejected utterances
how many were confirmation questions	How many times no answer detected
how many got answer YES	How many errors
how many got answer NO	How many DTMF digits pressed

Figure 3: list of low-level per-dialogue measures automatically computed.

## 4.6 Clinical Trial Design

The clinical trial was designed as a two-arm controlled study. Standard randomization tables were provided to clinicians to be used when enrolling patients. At enrolment, each patient is assigned either to the treatment or to the control group. The assignment was then recorded in the electronic health record system, together with the outcomes of the initial visit. Patients in the treatment group at this stage are given the Homey toll-free telephone number to call and a randomly generated 6-digit personal access code. Figure 1 shows the balance in control and treatment groups (respectively 66 and 59 patients). Total patients enrolled in hospital 2 are 70, in hospital 3 were 55. Hospitals coded 2 and 3 are a major institutions of Tuscany and Lombardy respectively.

### 4.6.1 Demographics

Age distribution of the patients enrolled in the study is given in the figure 4 below. Most of the patients are in the age range 50-65. The population fraction breakdown with respect to hospital and sex is shown in figure 3.

### 4.6.2 Call collection

At the time this report was written, the data collection was still ongoing. Total 547 calls from actual patients were collected, plus considering the 301 records entered by the physicians in the occurrence of face-to-face visits. The calls were collected and analyzed according by means of an ad-hoc architecture, shown in figure 4, developed to re-analyze logs and data inserted in the electronic health record (EPR). Figure 1 shows an example of the EPR, accessed from the web to enter or retrieve values measured. The figure, in particular, shows self-reported data (as seen by the top-right telephone icon).



Figure 3

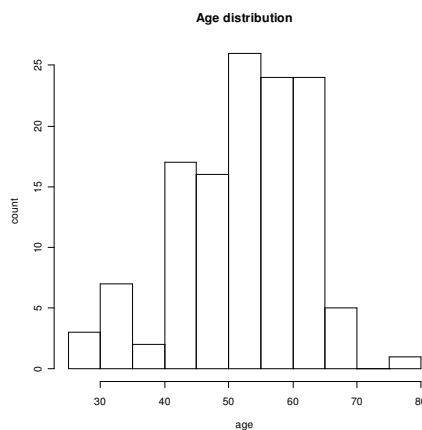


Figure 4

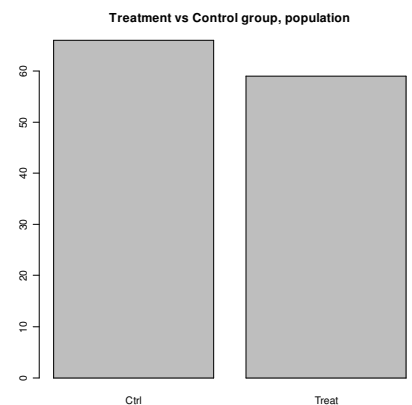


Figure 5

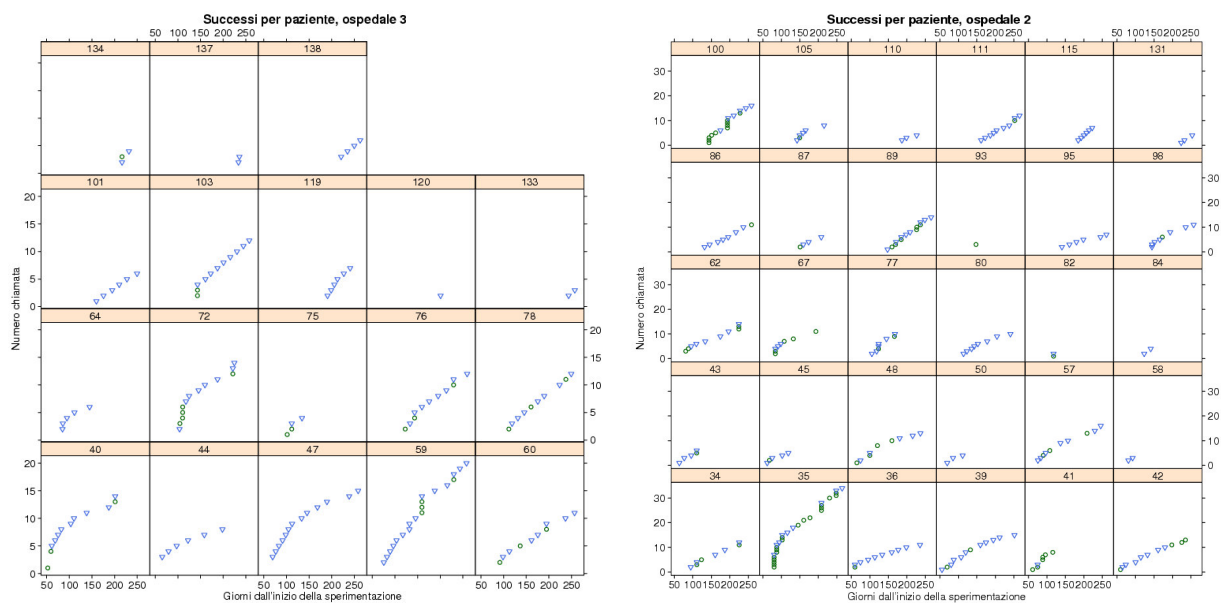
### 4.6.3 Call diagrams

After the internal evaluation was completed, the dialogue description was finalized and deployed for the clinical trial, which started on August 2003.

Figure 5 provides one view of the usage patterns of the telephone system, over time, for a subset of the treatment patients enrolled into the clinical trial. Each patient is shown as a Cartesian plane in a separate box. A dot is placed on the plane for each call; the  $x$  coordinate of the point relates to the day in which the call was received (day zero being the begin of the trial). The vertical axis displays the cumulative number of calls. Points have a different shape to indicate whether the dialogue reached the final salutation message (triangle) or not (small circle). This representation was chosen to distinguish conversations that ended properly from those which did so prematurely. Patient enrollment is still in progress, therefore not all “first calls” appear close to the date at the beginning of the trial.

Patterns in these plots can be recognized by inspection. Call diagrams like (a) correspond to people that regularly call the system and interact successfully with it. They show dominantly “good-end” calls. The slope of curves (a) and (b) is approximately one call every two weeks. Plot (b) shows that there has been a “training period”, during which patient 103 was unable to successfully complete two dialogues. The patient called back again the same day, and at the third attempt succeeded in entering the values; subsequent calls were performed smoothly and on schedule. Curve (c) shows that the respective patient followed the physician’s direction to call more frequently, once a week, during the first two months of the system’s use, then to reduce the number of calls to every other week.

People which call the system regularly, but experience usage problems, may show up as in diagrams (d) and (e). In the former, one finds a large fraction of “circle” symbols, but the curve does not have remarkable jumps: this means that on the average the call frequency was kept as required, despite of the outcome of specific calls. As a remark, the fact that a call is shown as a circle does not mean that the call was unsuccessful at acquiring values, but merely that the bye-message was not played: the more important clinical data were nevertheless recorded in the database. This is seen, for example, in (g), whose user from time to time



hangs up the conversation after the data acquisition phase. Plots like (e) show multiple attempts done in a row in the same day, or close dates: they appear as rising “steps” of terminated calls.

Patients that regrettably stop using the system altogether are also quite easily distinguished, as in (f). It is remarkable to note that not all quitters have a record of unsuccessful previous calls.

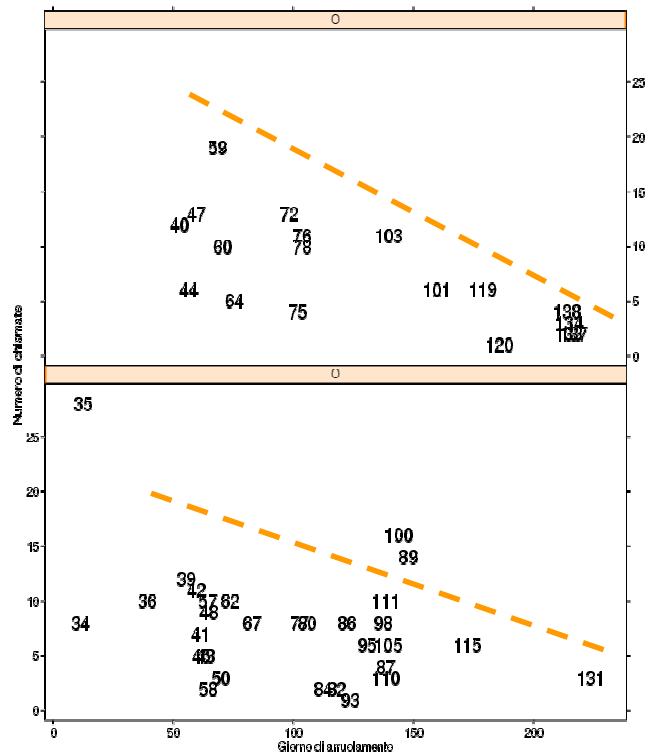
The call collection phase and the patient enrolment started at the same time, i.e. calls started coming in even when the enrolment process was still ongoing. This choice was made, given the length of the enrolment process, to minimize the time to complete the study.

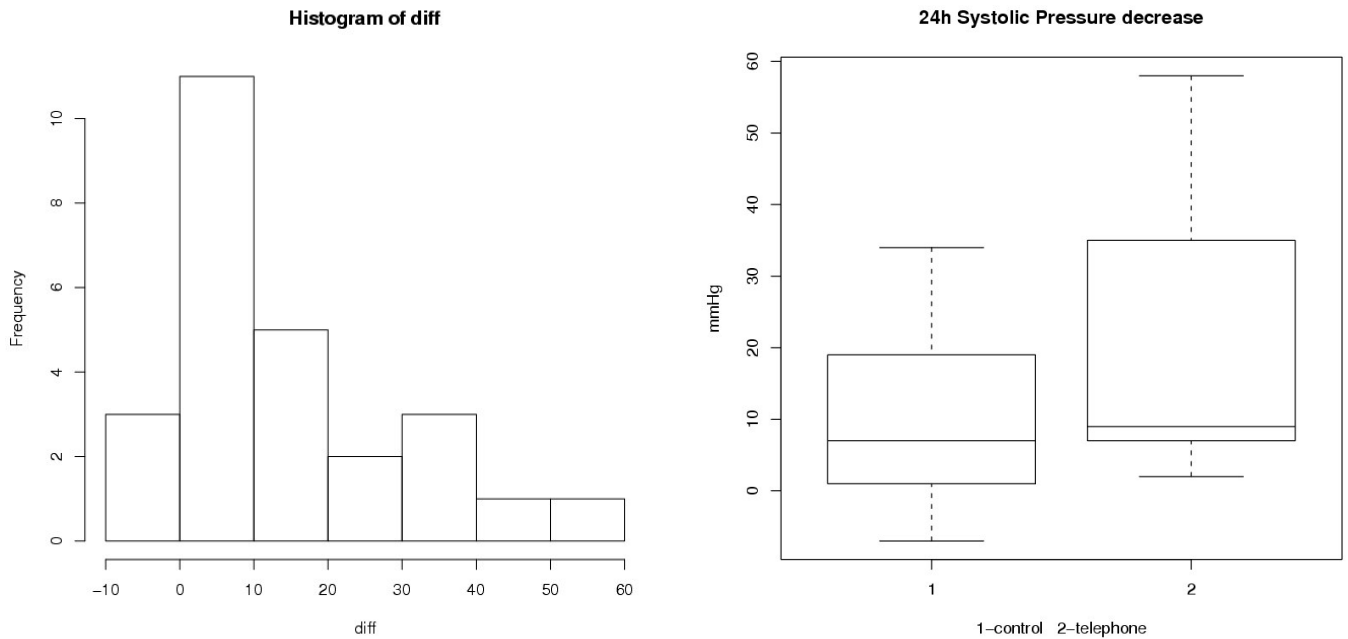
#### 4.6.4 Procedure

According to recent evidence, we chose to evaluate the clinical outcomes according to the *24-hour average blood pressure* (ABPM). ABPM samples patient’s blood pressure at regular time intervals over 24 hours (including day and night). The recordings are stored inside the apparatus and then transferred by the physician to the database. Patients approximately undergo ABPM every 6 months. ABPM is considered an accurate and objective measurement with respect to (1) self-reported blood pressure values, which has been shown to be affected by reporting biases, and (2) one-time ambulatory measurements, which are affected by the white-coat effect. ABPM need to be employed because the biases mentioned are typically of the same order of magnitude with respect to the decreasing in blood pressure that one hopes to detect in the study. At the time when this report was prepared, only a limited fraction of the patients has undergone at least two ABPM measurements. Two measurements, at a time distance apart, are necessary to assess the decreasing in blood pressure. The limited sample size has an impact on the statistical test’s p-values.

The statistical procedure followed for obtaining the results is as follows:

1. Patients with at least one ABPM measurement are selected





2. Per-patient ABPM measurements are arranged into couples. Members of the couple are time-ordered (older, newer measurement).
3. Each couple is replaced by the difference of its members
4. Statistical tests are performed on the vector of differences

#### 4.6.5 Results

Taking control and treatment patients together, there is *strong* statistical evidence that the 24-h average systolic blood pressure decreases. The p-value for the Wilcoxon nonparametric test is approximately  $10^{-4}$ . The histogram of blood pressure decrements is shown in the figure below, left. A clear peak shows that the majority of the patients enjoyed a systolic BP decrease between 0 to 10, and several had 10 to 20.

When the data difference set is classified (control vs treatment), there is *statistical hint* that the treatment group has a greater systolic pressure decrease with respect to the control group. Blood pressure decrements are given in a box-plot on the right hand side of the figure. Columns in the figure show decrement quartiles for the control group, and treatment population (columns 1 and 2 respectively). *Average* blood pressure decrease for the control group is 9.6; for the treatment is 19.6. Wilcoxon nonparametric test for non-equality in means yields a p-value for this conclusion of 0.1. Further investigation with more data is therefore necessary.

#### 4.7 Dialogue Refinements

In previous works (see falavigna1998, falavigna2000), ITC-Irst has developed a frame-based, mixed initiative dialogue system for telephone applications. During the HOMEY project (see Homey Deliverable 7) this system has been integrated in the telephone platform developed by Reitek, and a dialogue description for monitoring chronic patients affected by hypertension disease was designed. This dialogue model is capable of adapting itself to patient profiles (see Homey Deliverable 7, for more details). The telephone platform was successively installed in the CBIM labs, and a set of telephone calls, performed by patients engaged by two Italian hospitals (the Hospital of Pavia and the Hospital of Florence), have been recorded. These calls (each one corresponding to a dialogue) are still under way) and are periodically

checked and manually transcribed. After some months from the beginning of the calls, an evaluation of the data collected was done, which led to some minor modifications both in the dialogue strategy and in the recognition grammars.

#### 4.7.1 Problems encountered

The observed problems basically concern the following issues:

- *grammar coverage*, i.e. the percentage of sentences contained in the speech recognition grammars. Given a set of initial grammars, *grammar coverage* can be partially improved by refining the grammars themselves, following indications coming from a careful analysis of a certain amount of recorded and transcribed sentences. A sparseness problem arises due to the fact that at each dialogue state a different grammar is activated; this problem is mitigated by the usage of basic sub-grammars (e.g. for recognizing numbers, confirmations, dates, etc.) common to several dialogue states. Anyway, for some rarely used dialogue states, especially if the prompt does not lead to a nearly closed set of possible answers, grammar coverage still remains an open problem.
- *speech recognition performance*, in particular for recognizing the patient ages. We observed many substitution errors between the Italian numbers "sessanta" and "settanta" (60 and 70), even inside composite numbers as "centosessantaquattro" (164). This is due to the fact that the two words are acoustically similar, as they differ only for the phone /s:/ in 60 which becomes /t:/ in 70 (SAMPA units). The problem is enhanced by the fact that the telephone line cuts frequencies over 3600 Hz, i.e. in a frequency band that contains significant information for discriminating between phones /s/ and /t/. In some cases, patients tried to correct data by spelling numbers, e.g. "sessantasette sei sette" (67 6 7).
- *speech synthesis* suffered from the same problem. Sometimes, we observed that the patient did not confirm a value because he/she misunderstood the confirmation prompt (e.g. "minimal blood pressure 78, is it correct?"). This turns out to be a well-known problem, at least for people playing a game called "lotto" (a sort of bingo); when professional speakers read, in the TV, the extracted values they also spell the numbers (i.e. "34, 12, 67 6 7, 85, 70 7 0"), in order to avoid possible misunderstandings.
- *multiple confirmations* of basic data (e.g. min/max blood pressure, weight, cardiac frequency) was not totally effective with real users, despite the fact that during the first internal trial this seemed to be a better choice with respect to normal confirmation strategy. In fact, when the recognizer performs well, less turns are needed to complete this stage (6 vs. 8, see example below), as well as in the case of multiple errors (14 vs. 16). Of course, in both cases the number of turns can be reduced if immediate correction is performed. However, we noticed that some of the real users found pretty difficult to correctly answer to the question "tell me which data are wrong". Moreover, in case of multiple errors, sometimes they corrected only one of them, leaving unchanged (wrong) the other ones. These situations were never observed during the initial internal trials with volunteers.
- *minor dialogue refinements*, like unnecessary confirmations in some rarely used sub-dialogues.

<b>multiple confirmation strategy without errors (6 turns):</b>
S: please tell me min/max blood pressure
U: 67, 120
S: please tell me weight and cardiac frequency
U: 84, 70
S: min blood pressure 67, max blood pressure 120, weight 84, cardiac frequency 70?
U: yes

<b>normal confirmation strategy without errors (8 turns):</b>
S: please tell me min/max blood pressure
U: 67, 120
S: min blood pressure 67, max blood pressure 120?
U: yes
S: please tell me weight and cardiac frequency
U: 84, 70
S: weight 84, cardiac frequency 70?
U: yes

#### 4.7.2 Corrections

To overcome the problems described above, the following modifications, estimated on a set of manually checked and transcribed telephone calls, were applied to the dialogue:

- minor modifications to both grammars and dialogue flow, to improve coverage and to avoid unnecessary confirmations.
- grammars for recognizing numbers were modified to allow for spelled modality, in different combinations. This means that each number (e.g. 123) can be pronounced equivalently as:

(NUM3( 123 )NUM3)  
 (NUMDIG( 1 2 3 )NUMDIG)  
 (NUM3( 123 )NUM3) (NUMDIG( 1 2 3 )NUMDIG)  
 (NUMDIG( 1 2 3 )NUMDIG) (NUM3( 123 )NUM3)

Where **NUM3** is a grammar designed for recognizing numbers formed by three digits (i.e. one hundred and twenty three), and **NUMDIG** allows to recognize sequences of single digits (i.e. one two three). This modification leads to a better grammar coverage (hence, sentences previously impossible to being properly recognized now could be correctly recognized) whilst in-grammar sentences are worse recognized.

- multiple confirmation was replaced by immediate explicit confirmation.

When these corrections were activated, a prompt message informed the patients of the new system's capabilities.

### 4.7.3 Evaluation

Some weeks after the corrections, a subjective evaluation of the new dialogues was performed to verify their effectiveness. In particular we observed that:

- the normal confirmation strategy worked well;
- dialogues looked more fluent, due to some minor changes.
- spelled numbers were particularly appreciated. At the beginning, most of the patients tried this new function when entering numbers, and after a while they used it only for corrections. Considering that numbers can now be entered as either normal *numbers* (*NUM3( sessantasette )NUM3*) or in spelled modality (*NUMDIG( sei sette )NUMDIG*), in the new dialogues the grammar NUM3 was used 456 times, while NUMDIG was used 116 times. More specifically, the following statistics in terms of complete semantic sentences were observed:

346	(ITEM( (NUM3) )ITEM)
108	(ITEM( (NUMDIG) )ITEM)
49	(ITEM( (NUM3) )ITEM) (ITEM( (NUM3) )ITEM)
6	(YESNO( )YESNO) (ITEM( (NUM3) )ITEM)
3	(YESNO( )YESNO) (ITEM( (NUMDIG) )ITEM)
3	(ITEM( (NUMDIG) (NUM3) )ITEM)
2	(ITEM( (NUM3) (NUMDIG) )ITEM)

showing that only few times patients used a mixed modality.

- both coverage and recognition rate improved significantly (see section 4.9 for details).

Futhermore, after the dialogue corrections, the number of non completed dialogues dropped dramatically.

## 4.8 Data Collected

As seen above, all of the telephone calls were recorded. Each call corresponds to a patient dialogue, where the patient can be identified after the confirmation of a 6 digit identification code. For each dialogue, a complete log is available, including TTS prompts, speech signals, ASR output, timings, etc. Dialogues can be divided into 3 sets:

- **test patients**, performed by CBIM researchers which used their own (false) codes;
- **true patients**, performed by chronic patients correctly identified by their code;
- **anonymous patients**, when the system failed to recognize a valid code, often because long pauses among digits triggered the end point detector, and the patient preferred to hung up and call again rather than trying to correct the code.

All of the dialogues of true patients until September 2004 have been orthographically transcribed. As data collection is currently going on, all the results have to be considered as preliminary. So far, three test sets have been defined:

- **feb04** data: dialogues collected between August 2003 and February 2004. They amount to 5485 speech utterances, divided as

patients <b>feb04</b>	speakers	utterances	dialogues
test	11	1494	113
<b>true</b>	<b>44</b>	<b>3784</b>	<b>235</b>
anonymous (incomplete code)	20	207	29
<b>tot</b>	<b>75</b>	<b>5485</b>	<b>377</b>

True patients can be further divided according to their gender:

patients <b>feb04</b>	speakers	utterances	dialogues
female	19	1859	118
male	25	1925	117

- **aug04** data: dialogues collected after the dialogue refinements (end May 2004) between June 2004 and August 2004. They contain 2277 sentences, and were mainly used to verify the effectiveness of the modifications described above.
- **sep04** data: dialogues collected between august 2003 and September 2004. They include both feb04 and aug04 data and contain 9527 speech utterances, for a total of 541 dialogues, 54 speakers.

true patients <b>sep04</b>	speakers	utterances	dialogues
female	21	4242	244
male	33	5285	297
<b>tot</b>	<b>54</b>	<b>9527</b>	<b>541</b>

## 4.9 Results

### 4.9.1 Grammar Coverage

Grammar coverage is evaluated by summing the coverage of each grammar used in the dialogue description. The number of different grammars was 43 before the modification, 36 after the modifications. For each grammar, two recognition lists can be defined: all of the audio files acquired in a certain dialogue state, and the subset of the sentences covered by that grammar. Improving coverage normally leads to a trade-off: better recognition rate on the whole list, worse recognition rate on the covered subset. By summing up the contributions for each grammar, the following results in terms of coverage were obtained. Note that for the set04 set (\*) we used the last version of the grammars, thus only the sentences that were recognized with grammars which survived the dialogue/grammar refinement were considered.

data set	global coverage	sentences recognized/tot sentences
feb04	90.46%	( 3423 / 3784 )
aug04	95.87%	( 2183 / 2277 )
set04	94.18%	( 8722 / 9261 ) *

From this table it can be seen that the percentage of out-of-grammar sentences was significantly reduced after the refinement.

## 4.9.2 Speech Recogniser Performance

### *Word and Sentence Accuracy*

During the project, we had to deal with different test sets, defined in section 4.8 and different grammar versions. In the following, results are reported in chronological order.

The next table reports results on the whole feb04 test set, as well as results for the two gender dependent subsets. In the following tables, StringRR stands for String Recognition Rate (percentage of correctly recognized sentences), UnitRR stands for Unit Recognition Rate (percentage of correctly recognized units), Units is the number of units pronounced in the data set, Errs report in detail the number of Deletions, Insertions, Substitutions. In the following, units are words. Female patients perform significantly worse than male ones.

test set	StringRR	UnitRR	Units	Errs (D+I+S)
feb04	77.85%	72.77%	6218	1693 (314+394+985)
feb04-female	73.00%	68.24%	3193	1014 (167+229+618)
feb04-male	82.55%	77.55%	3025	679 (147+165+367)

The following table reports the most frequent errors found in feb04: note that some of them will not result in dialogue problems, as they are insertions/deletions of semantically irrelevant words or substitutions which do not alter the semantic of the sentence (questa → questo, mattina → mattino).

SUBSTITUTIONS		INSERTIONS	DELETIONS		
21	questa → questo	38	ho	24	si'
15	si' → certo	30	un	18	sessanta
14	settantotto → ottantotto	25	e	17	uno
14	mattina → mattino	22	otto	14	sei
12	sessantacinque → settantacinque	22	di	13	due
9	tre → sei	20	una	12	non
9	sei → settantasei	9	pressione	9	settanta
8	sessantasei → settantasei	9	non	7	il
7	sessanta → settanta	7	tre	7	e
7	otto → ho	5	so	6	sette
6	uno → otto	5	sette	6	lo

The refinement of the dialogue led not only to a better grammar coverage (see section 4.9.1), but also to an improvement in speech recognition rate, as the following table shows. Note that the comparison is not straightforward, because the test set are different. Moreover, other causes (for instance, the expertise level of the patients) may affect results. Anyway, results in terms of word accuracy show a significant improvement.

test set	StringRR	UnitRR	Units	Errs (D+I+S)
feb04	77.85%	72.77%	6218	1693 (314+394+985)
aug04	88.98 %	88.82 %	3488	390 (164+56+170)

Performance further increase if computed on the subset of sentences having grammar coverage, both in feb04 and in aug04:

test set	StringRR	UnitRR	Units	Errs (D+I+S)
feb04-cov	85.25%	85.79%	5195	738 (88+204+446)
aug04-cov	91.20 %	92.30 %	3287	253 (85+40+128)

The next figures are related to the whole test set, sep04. The grammars used to recognize the sentences are the final ones. Speech recognition rate on the sep04 set, computed using the last version of the grammars, shows the same behaviour as in the previous cases.

test set	StringRR	UnitRR	Units	Errs (D+I+S)
sep04-cov	90.09 %	91.56 %	12908	1090 (271+172+647)
sep04	85.78 %	84.97 %	14344	2156 (879+317+960)
sep04-female	84.33 %	83.56 %	6726	1106 (407+162+537)
sep04-male	87.00 %	86.22 %	7618	1050 (472+155+423)

In the last figure, dialogues have been grouped to highlight how user expertise impacts on recognition rate. The first row groups each first and second dialogue of each patient; the second row groups dialogues from number 3 to number 4, and so on. The size of the data decreases as the dialogue number increase (only a few patients reached dialogue number 30); nevertheless there is a clear improvement in recognition rate.

test set	StringRR	UnitRR	Units	Errs (D+I+S)
dial01-02	83.30 %	79.94 %	2802	562 (283+76+203)
dial03-04	83.65 %	83.00 %	2471	420 (163+60+197)
dial05-06	88.14 %	87.84 %	2213	269 (94+37+138)
dial07-08	86.63 %	84.11 %	1693	269 (123+37+109)
dial09-10	85.78 %	85.40 %	1301	190 (66+34+90)
dial11-12	87.24 %	89.25 %	995	107 (32+19+56)
dial13-14	86.70 %	87.26 %	683	87 (29+18+40)
dial15-16	88.95 %	90.94 %	563	51 (16+ 9+26)
dial17-18	85.79 %	86.97 %	614	80 (28+11+41)
dial19-20	88.26 %	88.40 %	362	42 (21+ 6+15)
dial21-22	87.27 %	87.36 %	174	22 (11+ 3+ 8)
dial23-24	88.68 %	91.89 %	74	6 ( 1+ 1+ 4)
dial25-26	71.43 %	75.61 %	41	10 ( 1+ 2+ 7)
dial27-28	68.09 %	71.21 %	66	19 ( 1+ 2+16)
dial29-30	93.55 %	95.35 %	43	2 ( 0+ 1+ 1)
dial31-32	82.22 %	86.67 %	60	8 ( 4+ 0+ 4)
dial33-34	95.24 %	95.24 %	63	3 ( 2+ 0+ 1)
dial35-36	95.35 %	95.24 %	63	3 ( 2+ 0+ 1)
dial37-38	86.36 %	90.48 %	63	6 ( 2+ 1+ 3)

### *Semantic Accuracy*

In a dialogue application it makes sense to compute semantic accuracy, which better represent the dialogue behaviour than word accuracy. However, the only available transcription was orthographic, so we had to automatically produce a semantic representation by using the recognizer in text mode. Occasionally, in case of ambiguous sentences, this process can result in an error in the semantic representation. As before, in the following table, StringRR stands for String Recognition Rate (percentage of correctly recognized sentences),

UnitRR stands for Unit Recognition Rate (percentage of correctly recognized units), Units is the number of units pronounced in the data set, Errs report in detail the number of Deletions, Insertions, Substitutions. In the following table, units are no longer words, but semantic concepts.

test set	StringRR	UnitRR	Units	Errs (D+I+S)
sep04	88.10 %	87.40 %	9239	1164 (228+367+569)
sep04-cov	91.05 %	91.27 %	9235	806 (227+12+567)

## 5 Validation of the Multimodal Browser

As reported in the technical annex of the Homey project, one of the most challenging objectives of the project has been to develop a system architecture capable of handling different input/output interacting modalities. Deliverables 4 and 5 give the details of such architecture, including the multi-modal markup language specification and the alignment/synchronization protocols used in the developed prototypes. In September 2004 the multimodal prototype, developed within the Homey project, has been validated over the Electronic Patient Record (EPR), developed by the Telemedicine Division of ITC-Irst for the oncology domain (see Galligioni2001, for the details). Furthermore, the prototype has been used inside the oncological clinic of the S. Chiara Hospital in Trento (Italy), for entering (via automatic speech recognition) the laboratory test results brought by the patients of the clinic itself during their day hospitals. This task is mainly a heavy data entry task, in which reception nurses have to go through patient information and entry many numerical values related to laboratory test results.

The system has been used by both nurses of the clinic and by some researchers of the Telemedicine Division of ITC-Irst, and comparisons have been made between traditional data entry (i.e. through mouse and keyboard) and multimodal data entry, carried out by both keyboard and automatic speech recognition. Results are given below.

In a first step, we evaluated the efficacy of voice interaction in data entry task by measuring accuracy and completion time. A total of 8 subjects participated in the experiment: five nurses of the oncological ward who have routinely used the EMR for 3 years and have a significant experience with data entry by keyboard. Two of them (nurse A and nurse B) demonstrate a better skill in using computers. Three researchers of the Telemedicine Division of ITC-Irst have also used the system: none of them were familiar with speech recognition and no training activity was performed.

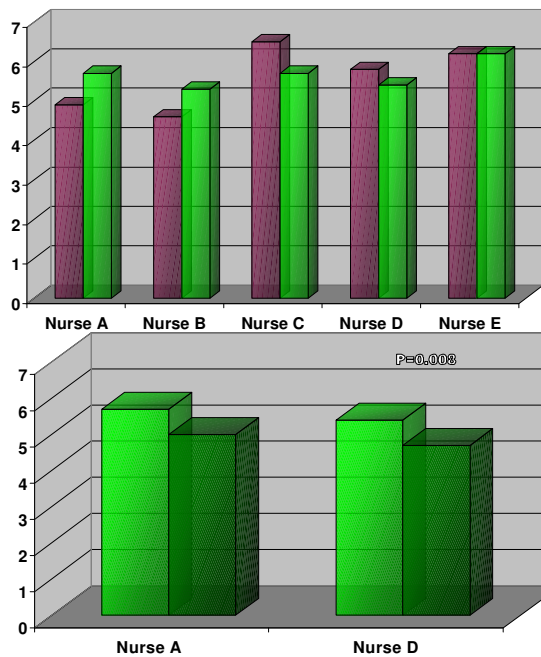
The experiment was performed in three different sessions. All participants had to insert a total of 13 clinical test results in the traditional manner via keyboard (6 regarding biochemical and 7 regarding blood parameters with 12 and 5 fields respectively).

The same values were inserted by the same participants using the voice. Users uttered the name of the field followed by its value and then controlled the correctness of the entries. If errors occurred the wrong fields had to be re-uttered.

Two nurses inserted the clinical test results by voice a second time after a week. The goal of the 1st and 2nd session was to estimate the performance of the data entry task by keyboard and by voice respectively. The goal of the 3rd one was to have a preliminary estimation of the learning curve of the voice tool. The obtained performance are reported in the graphs of figures 2, 3, 4, where the colours of the vertical bars correspond to either voice and keyboard data entry, as well as to the first and second acquisition phases as given in the legend of figure 5.

The average word accuracy measured on the collected data (note that the speech recognition task mainly consists of decimal numbers) was 91%, ranging from 86%, for the worst speaker, to 100% for the best one.

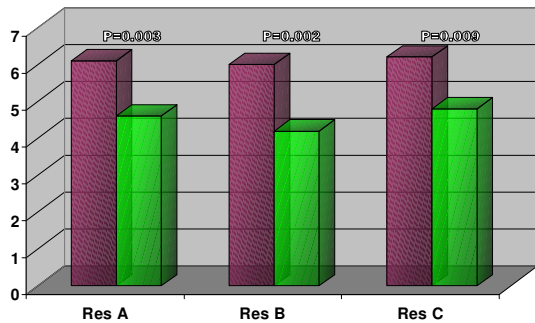
Figure 2: average data entry time required by nurses.



For every subject, the difference between voice and keyboard entry time is not statistically significant. Furthermore, the average time the nurses with a skill using computer (nurse A and B) spent inserting the data by pressing keys is significantly less than the time spent uttering the fields ( $P=0.005$ ). The average time the other nurses spent using the keyboard is greater than the time spent using the voice

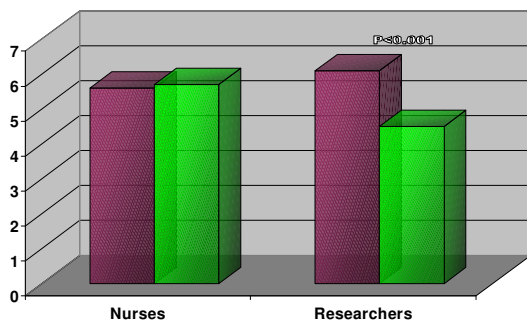
There is an improvement, from the 1st to the 2nd session, of the voice data entry speed of the two nurses (an average gain of 0.7 sec for both) but only for the nurse D is statistically significant ( $P=0.008$ ).

Figure 3: average data entry time required by researchers.



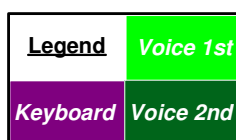
There is a statistically significant improvement of the researchers' performance of data entry by voice respect to keyboard ( $P=0.001$ ). The average gain is about 1.5 sec for entry.

Figure 4: comparison of data entry times between nurses and researchers.



The nurses and the researchers have no significant different data entry times using the keyboard, whilst, using the voice, researchers spent less time than nurses ( $P<0.001$ ).

Figure 5: legend for figures 2, 3 and 5.



From this preliminary evaluation phase, we can conclude that speech recognition seems to be an acceptable alternative to conventional data entry task. The main advantage of using voice may be that it requires a shorter learning curve.

## 6 References

- Stefanelli, M., Rognoni, C., Quaglini, S., Giorgino, T., Piazza, M., and Azzini, I. (2003). Data Collection and Refinement of the Algorithms. Deliverable D7, HOMEY Project, IST-2001-32434.
- Berman, A., Cooper, R., Ericsson, S., Hieronymus, J., Jonson, R., Larsson, S., Milward, D., and Torre, D. (2000). *Implemented SIRIDUS System Architecture (Baseline)*. Deliverable D6.2, SIRIDUS Project, IST-1999-10516.
- Beveridge, M. A., and Milward, D. (2003). The High-Level Task Specification Language. Deliverable D11, HOMEY Project, IST-2001-32434.
- Bohlin, P., Bos, J., Larsson, S., Lewin, I., Matheson, C. & Milward D. (1999). *Survey of Existing Interactive Systems*. Deliverable D1.3, TRINDI Project, LE4-8314.
- Bury, J., Humber, M., and Fox, J. (2001). Integrating Decision Support with Electronic Referrals. In R. Rogers, R. Haux and V. Patel (Eds) *Medinfo*. IOS Press, Amsterdam.
- Ceusters, W., Beveridge, M. A., Milward, D., and Falavigna, D. (2002). *Specification for Semantic Dictionary Integration*, Deliverable D9, HOMEY Project, IST-2001-32434.
- De Mori R, et al. (1998) *Spoken Dialogues with Computers*, Academic Press, 1998.
- Emery J., Walton R., Murphy M., Austoker J., Yudkin P., Chapman C., Coulson A., Glasspool D., Fox J. (2000) Computer Support for Recording and Interpreting Family Histories of Breast and Ovarian Cancer in Primary Care: Comparative Study with Simulated Cases. *British Medical Journal* 321 pp. 28-32.
- Falavigna D., Gretter R.: Telephone Speech Recognition Applications at IRST, in Proceedings of 4th IEEE workshop on Interactive Voice Technology for Telecommunications Applications", Turin, Italy, September 1998
- Falavigna D., Gretter R., Orlandi M.: "A Mixed Language Model for a Dialogue System over the Telephone", in proceedings of ICSLP 2000, Beijing, China, 16-20 October 2000 - IRST Tech. Rep. No. 0003-32.
- Fox, J., Beveridge, M. A., and Glasspool, D. (2003). Understanding Intelligent Agents: Analysis and Synthesis. *AI Communications*, 16 (3) pp. 139-152.
- Galligioni et al. (2001). A Teleconsulting network between peripheral hospitals and the referring center for cancer patients in Trento (Italy). In *European Journal of Cancer*, Vol. 37, Suppl. 6: S23
- Heid, U., Bernsen, N., and Dybkjaer, L. (1998). *Current Practice in the Development and Evaluation of Spoken Language Dialogue Systems*. Deliverable D1.8, DISC Project, Esprit Long-Term Research Concerted Action No. 24823.
- Simpson, A., and Fraser (1993). Black Box and Glass Box Evaluation of the SUNDIAL System. In Proc. 3<sup>rd</sup> European Conference on Speech Communication and Technology (Eurospeech'93), Berlin, Germany.
- Stefanelli, M., Rognoni, C., Quaglini, S., Giorgino, T., and Azzini, I. (2002). Algorithms and Tools for Dialogue Adaptation. Deliverable D6, HOMEY Project, IST-2001-32434.
- Stefanelli, M., Rognoni, C., Quaglini, S., Giorgino, T., Piazza, M., and Azzini, I. (2003). Data Collection and Refinement of the Algorithms. Deliverable D7, HOMEY Project, IST-2001-32434.
- Sutton, S., Hansen, B., Lander, T., Novick, D. G., and Cole, R. (1995). Evaluating the Effectiveness of Dialogue for an Automated Spoken Questionnaire. Technical Report CS/E95-12, Dept. of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology.
- Walker, M. A., Litman, D. J., Candace, A. K., and Abella, A. (1998). Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies, *Computer Speech and Language* 12 (3)

Walker, M. A., Hirschman, L., and Aberdeen, J. (2000). Evaluation for DARPA Communicator Spoken Dialogue Systems, Proc. of 2<sup>nd</sup> International Conference on Language Resources and Evaluation, Athens, Greece.

## Appendix A: Script for Cancer Showcase Evaluation

S: Hello, welcome to the breast cancer referrals system.

U: **Hello**

S: What is the patient's age?

U: **She is thirty, female, and she has a bilateral nipple discharge**

S: Ok

S: And is it a bloodstained nipple discharge?

U: **No**

S: Ok

S: And is it a cloth-staining nipple discharge?

U: **No**

S: Ok

S: Does the patient have an acquired nipple deformity?

U: **What do you mean?**

S: For example: gestational inversion of nipple or gestational nipple retraction

U: **No, and she doesn't have a lump, or any eczema**

S: Ok

S: Does the patient have a skin ulcer?

U: **Yes**

S: Ok

S: Does the patient have any intractable pain?

U: **Yes**

S: Ok

S: Does the patient have a breast abscess?

U: **Yes, and she has a cyst**

S: Ok

S: Do you mean a breast cyst?

U: **Yes**

S: Ok

S: So this patient is thirty, female, and has a skin ulcer, intractable pain, a breast abscess, a breast cyst, and a bilateral nipple discharge. Is that correct?

U: **No, she isn't thirty<sup>3</sup>**

S: What is the patient's age?

U: **Thirty-five**

S: Ok

---

<sup>3</sup> or whatever number the system says...

S: So this patient is thirty-five, female, and has a skin ulcer, intractable pain, a breast abscess, a breast cyst, and a bilateral nipple discharge. Is that correct?

U: **Yes**

S: I recommend urgent referral to breast specialist. Which action do you advise: urgent referral to breast specialist, support by general practitioner or non-urgent referral to breast specialist?

U: **Why do you recommend urgent referral?**

S: Urgent referral to breast specialist is indicated by the fact that this patient has a skin ulcer.

U: **What are the arguments for non-urgent referral**

S: Non-urgent referral to breast specialist is indicated by the fact that this patient has intractable pain and a breast abscess.

U: **Are there any arguments for support by general practitioner?**

S: There are no arguments for support by general practitioner.

U: **Ok, I advise urgent referral**

S: Ok

S: Do you mean urgent referral to breast specialist?

U: **Yes**

S: Ok

S: An urgent referral to breast specialist has been made for this patient.

U: **Ok**

S: Thankyou for using this service. Goodbye.

U: **Bye**