A Details of the Algorithm

A.1 Forward-Backward Procedure

We compute the necessary marginalizations of the joint distribution $\mathbb{P}(\bar{\mathcal{D}}|\mathcal{D}, \hat{\theta})$ using the forward-backward algorithm. Letting $\boldsymbol{x}_{t:t'} = \{x_t, x_{t+1}, \ldots, x_{t'}\}$ for any time-indexed quantity x_t, the forward messages are defined as $\alpha_t(s) = \mathbb{P}(s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}|\hat{\theta})$, which can be computed dynamically as

$$\begin{aligned}
\alpha_{t+1}(s') &= \mathbb{P}(s_{t+1} = s', \boldsymbol{a}_{1:t}, \boldsymbol{z}_{1:t}|\hat{\theta}) \\
&= \sum_{s \in S} \mathbb{P}(s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}|\hat{\theta}) \mathbb{P}(s_{t+1} = s', a_t, z_t | s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}, \hat{\theta}) \\
&= \sum_{s \in S} \alpha_t(s) \hat{\pi}(a_t|b_t) \hat{T}(s'|s, a_t) \hat{O}(z_t|a_t, s') \\
&\propto \sum_{s \in S} \alpha_t(s) \hat{T}(s'|s, a_t) \hat{O}(z_t|a_t, s')
\end{aligned}$$

with initial case $\alpha_1(s) = \mathbb{P}(s_1 = s) = b_1(s)$. The backward messages are defined as $\beta_t(s) = \mathbb{P}(\boldsymbol{a}_{t:\tau}, \boldsymbol{z}_{t:\tau}|s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}, \hat{\theta})$, which can also be computed dynamically as

$$\begin{aligned}
\beta_t(s) &= \mathbb{P}(\boldsymbol{a}_{t:\tau}, \boldsymbol{z}_{t:\tau}|s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}, \hat{\theta}) \\
&= \sum_{s' \in S} \mathbb{P}(s_{t+1} = s', a_t, z_t | s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}, \hat{\theta}) \mathbb{P}(\boldsymbol{a}_{t+1:\tau}, \boldsymbol{z}_{t+1:\tau}|s_{t+1} = s', \boldsymbol{a}_{1:t}, \boldsymbol{z}_{1:t}, \hat{\theta}) \\
&= \sum_{s' \in S} \hat{\pi}(a_t|b_t) \hat{T}(s'|s, a_t) \hat{O}(z_t|a_t, s') \beta_{t+1}(s') \\
&\propto \sum_{s' \in S} \hat{T}(s'|s, a_t) \hat{O}(z_t|a_t, s') \beta_{t+1}(s')
\end{aligned}$$

with initial case $\beta_{\tau+1}(s) = \mathbb{P}(\emptyset|s_{\tau+1} = s, \boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}, \hat{\theta}) = 1$.

Then, the marginal probability of being in state s at time t given the dataset $\mathcal{D}$ and the estimate $\hat{\theta}$ can be computed as

$$\begin{aligned}
\gamma_t(s) &= \mathbb{P}(s_t = s|\mathcal{D}, \hat{\theta}) \\
&= \mathbb{P}(s_t = s|\boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}, \hat{\theta}) \\
&\propto \mathbb{P}(s_t = s, \boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}|\hat{\theta}) \\
&= \alpha_t(s)\beta(s)
\end{aligned}$$

and similarly, the marginal probability of transitioning from state s to state s' at the end of time t given the dataset $\mathcal{D}$ and the estimate $\hat{\theta}$ can be computed as

$$\begin{aligned}
\xi_t(s, s') &= \mathbb{P}(s_t = s, s_{t+1} = s'|\mathcal{D}, \hat{\theta}) \\
&\propto \mathbb{P}(s_t = s, s_{t+1} = s', \boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}|\hat{\theta}) \\
&= \mathbb{P}(s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}|\hat{\theta}) \mathbb{P}(s_{t+1} = s', a_t, z_t | s_t = s, \boldsymbol{a}_{1:t-1}, \boldsymbol{z}_{1:t-1}, \hat{\theta}) \\
&\qquad \times \mathbb{P}(\boldsymbol{a}_{t+1:\tau}, \boldsymbol{z}_{t+1:\tau}|s_{t+1} = s', \boldsymbol{a}_{1:t}, \boldsymbol{z}_{1:t}, \hat{\theta}) \\
&= \alpha_t(s) \hat{\pi}(a_t|b_t) \hat{T}(s'|s, a) \hat{O}(z|a, s') \beta_{t+1}(s') \\
&\propto \alpha_t(s) \hat{T}(s'|s, a) \hat{O}(z|a, s') \beta_{t+1}(s') \, .
\end{aligned}$$

A.2 Gradient-Ascent Procedure

Taking the gradient of the expected log-likelihood $Q(\theta; \hat{\theta})$ in (5) with respect to the unknown parameters $\theta = (T, O, b_1, \eta, \mu_{a \in A})$ first requires computing the Jacobian matrix $\nabla_{b_t} b_{t'}$ for $1 \leq$

$t < t' \leq \tau$, where $(\nabla_b b')_{ij} = \partial b'(i)/\partial b(j)$ for $i, j \in S$. This can be achieved dynamically as $\nabla_{b_t} b_{t'} = \nabla_{b_{t+1}} b_{t'} \nabla_{b_t} b_{t+1}$ with initial case $\nabla_{b_{t'}} b_{t'} = I$, where

$$\begin{aligned}
(\nabla_{b_t} b_{t+1})_{ij} &= \frac{\partial b_{t+1}(i)}{\partial b_t(j)} \\
&= \frac{\partial}{\partial b_t(j)} \left[\frac{\sum_{x \in S} b_t(x) T(i|x, a_t) O(z_t|a_t, i)}{\sum_{x \in S} \sum_{x' \in S} b_t(x) T(x'|x, a_t) O(z_t|a_t, x')} \right] \\
&= \frac{T(i|j, a_t) O(z_t|a_t, i)}{\sum_{x \in S} \sum_{x' \in S} b_t(x) T(x'|x, a_t) O(z_t|a_t, x')} \\
&\qquad - \frac{\sum_{x' \in S} T(x'|j, a_t) O(z_t|a_t, x')}{(\sum_{x \in S} \sum_{x' \in S} b_t(x) T(x'|x, a_t) O(z_t|a_t, x'))^2} \, .
\end{aligned}$$

A.2.1 Partial Derivatives

The derivative of $Q(\theta; \hat{\theta})$ with respect to $T(s'|s, a)$ is

$$\begin{aligned}
\frac{\partial Q(\theta; \hat{\theta})}{\partial T(s'|s, a)} &= \frac{\partial}{\partial T(s'|s, a)} \sum_{i=1}^{n} \left[\sum_{t=1}^{\tau} \mathbb{I}\{a_t = a\} \sum_{x \in S} \sum_{x' \in S} \xi_t(x, x') \log T(x'|x, a) \right. \\
&\qquad \left. + \sum_{t=2}^{\tau} \log \pi(a_t|b_t) \right] \\
&= \sum_{i=1}^{n} \left[\sum_{t=1}^{\tau} \mathbb{I}\{a_t = a\} \frac{\xi_t(s, s')}{T(s'|s, a)} + \sum_{t=2}^{\tau} \frac{\partial \log \pi(a_t|b_t)}{\partial T(s'|s, a)} \right] \\
&= \sum_{i=1}^{n} \left[\sum_{t=1}^{\tau} \mathbb{I}\{a_t = a\} \frac{\xi_t(s, s')}{T(s'|s, a)} \right. \\
&\qquad \left. + \sum_{t=2}^{\tau} \sum_{t'=1}^{t-1} \nabla_{b_t} \log \pi(a_t|b_t) \nabla_{b_{t'+1}} b_t \nabla_{T(s'|s,a)} b_{t'+1} \right] \, ,
\end{aligned}$$

where

$$\begin{aligned}
(\nabla_{b_t} \log \pi(a_t|b_t))_{1j} &= \frac{\partial \log(a_t|b_t)}{\partial b_t(j)} \\
&= \frac{\partial}{\partial b_t(j)} \left(-\eta \|b_t - \mu_{a_t}\|^2 - \log \sum_{a' \in A} e^{-\eta \|b_t - \mu_{a'}\|^2} \right) \\
&= -2\eta (b_t(j) - \mu_{a_t}(j)) + 2\eta \sum_{a \in A} \frac{e^{-\eta \|b_t - \mu_a\|^2}}{\sum_{a' \in A} e^{-\eta \|b_t - \mu_{a'}\|^2}} (b_t(j) - \mu_a(j)) \\
&= -2\eta (b_t(j) - \mu_{a_t}(j)) + 2\eta \sum_{a \in A} \pi(a|b_t)(b_t(j) - \mu_a(j))
\end{aligned}$$

and

$$\begin{aligned}
(\nabla_{T(s'|s,a)} b_{t'+1})_{i1} &= \frac{\partial b_{t'+1}(i)}{\partial T(s'|s, a)} \\
&= \frac{\partial}{\partial T(s'|s, a)} \left(\frac{\sum_{x \in S} b_{t'}(x) T(i|x, a_{t'}) O(z_{t'}|a_{t'}, i)}{\sum_{x \in S} \sum_{x' \in S} b_{t'}(x) T(x'|x, a_{t'}) O(z_{t'}|a_{t'}, x')} \right) \\
&= \mathbb{I}\{a_{t'} = a\} \left(\frac{\mathbb{I}\{i = s'\} b_{t'}(s) O(z_{t'}|a, s')}{\sum_{x \in S} \sum_{x' \in S} b_{t'}(x) T(x'|x, a) O(z_{t'}|a, x')} \right. \\
&\qquad \left. - \frac{b_{t'}(s) O(z_{t'}|a, s')}{(\sum_{x \in S} \sum_{x' \in S} b_{t'}(x) T(x'|x, a) O(z_{t'}|a, x'))^2} \right) \, .
\end{aligned}$$

The derivative of $Q(\theta; \hat{\theta})$ with respect to $O(z|a, s')$ is

$$\frac{\partial Q(\theta; \hat{\theta})}{\partial O(z|a, s')} = \frac{\partial}{\partial O(z|a, s')} \sum_{i=1}^{n} \left[\sum_{t=1}^{\tau} \mathbb{I}\{a_t = a, z_t = z\} \sum_{x' \in S} \gamma_{t+1}(x') \log O(z|a, x') \right.$$

$$
\begin{aligned}
&\qquad\qquad + \sum_{t=2}^{\tau} \log \pi(a_t|b_t) \Bigg] \\
&= \sum_{i=1}^{n} \left[\sum_{t=1}^{\tau} \mathbb{I}\{a_t = a, z_t = z\} \frac{\gamma_{t+1}(s')}{O(z|a,s')} + \sum_{t=2}^{\tau} \frac{\partial \log \pi(a_t|b_t)}{\partial O(z|a,s')} \right] \\
&= \sum_{i=1}^{n} \left[\sum_{t=1}^{\tau} \mathbb{I}\{a_t = a, z_t = z\} \frac{\gamma_{t+1}(s')}{O(z|a,s')} \right. \\
&\qquad\qquad \left. + \sum_{t=2}^{\tau} \sum_{t'=1}^{t-1} \nabla_{b_t} \log \pi(a_t|b_t) \nabla_{b_{t'+1}} b_t \nabla_{O(z|a,s')} b_{t'+1} \right] ,
\end{aligned}
$$

where

$$
\begin{aligned}
(\nabla_{O(z|a,s')} b_{t'+1})_{i1} &= \frac{\partial b_{t'+1}(i)}{\partial O(z|a,s')} \\
&= \frac{\partial}{\partial O(z|a,s')} \left(\frac{\sum_{x \in S} b_{t'}(x) T(i|x,a_{t'}) O(z_{t'}|a_{t'},i)}{\sum_{x \in S} \sum_{x' \in S} b_{t'}(x) T(x'|x,a_{t'}) O(z_{t'}|a_{t'},x')} \right) \\
&= \mathbb{I}\{a_{t'} = a, z_{t'} = z\} \left(\frac{\mathbb{I}\{i = s'\} \sum_{x \in S} b_{t'}(x) T(s'|x,a)}{\sum_{x \in S} \sum_{x' \in S} b_{t'}(x) T(x'|x,a) O(z|a,x')} \right. \\
&\qquad\qquad \left. - \frac{\sum_{x \in S} b_{t'}(x) T(s'|x,a)}{(\sum_{x \in S} \sum_{x' \in S} b_{t'}(x) T(x'|x,a) O(z|a,x'))^2} \right) .
\end{aligned}
$$

The derivative of $Q(\theta; \hat{\theta})$ with respect to $b_1(s)$ is

$$
\begin{aligned}
\frac{\partial Q(\theta;\hat{\theta})}{\partial b_1(s)} &= \frac{\partial}{\partial b_1(s)} \sum_{i=1}^{n} \left[\sum_{x \in S} \gamma_1(x) \log b_1(x) + \sum_{t=1}^{\tau} \log \pi(a_t|b_t) \right] \\
&= \sum_{i=1}^{n} \left[\frac{\gamma_1(s)}{b_1(s)} + \sum_{t=1}^{\tau} \nabla_{b_t} \log \pi(a_t|b_t) \nabla_{b_1(s)} b_t \right] ,
\end{aligned}
$$

where $(\nabla_{b_1(s)} b_t)_{i1} = (\nabla_{b_1} b_t)_{is}$.

The derivative of $Q(\theta; \hat{\theta})$ with respect to η is

$$
\begin{aligned}
\frac{\partial Q(\theta;\hat{\theta})}{\partial \eta} &= \frac{\partial}{\partial \eta} \sum_{i=1}^{n} \sum_{t=1}^{\tau} \log \pi(a_t|b_t) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{\tau} \frac{\partial}{\partial \eta} \left(-\eta \|b_t - \mu_{a_t}\|^2 - \log \sum_{a' \in A} e^{-\eta \|b_t - \mu_{a'}\|^2} \right) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{\tau} \left(-\|b_t - \mu_{a_t}\|^2 + \sum_{a \in A} \frac{e^{-\eta \|b_t - \mu_a\|^2}}{\sum_{a' \in A} e^{-\eta \|b_t - \mu_{a'}\|^2}} \|b_t - \mu_a\|^2 \right) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{\tau} \left(-\|b_t - \mu_{a_t}\|^2 + \sum_{a \in A} \pi(a|b_t) \|b_t - \mu_a\|^2 \right) .
\end{aligned}
$$

Finally, the derivative of $Q(\theta; \hat{\theta})$ with respect to $\mu_a(s)$ is

$$
\begin{aligned}
\frac{\partial Q(\theta;\hat{\theta})}{\partial \mu_a(s)} &= \frac{\partial}{\partial \mu_a(s)} \sum_{i=1}^{n} \sum_{t=1}^{\tau} \log \pi(a_t|b_t) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{\tau} \frac{\partial}{\partial \mu_a(s)} \left(-\eta \|b_t - \mu_{a_t}\|^2 - \log \sum_{a' \in A} e^{-\eta \|b_t - \mu_{a'}\|^2} \right) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{\tau} \left(2\eta \mathbb{I}\{a_t = a\} (b_t(s) - \mu_a(s)) - 2\eta \frac{e^{-\eta \|b_t - \mu_a\|^2}}{\sum_{a' \in A} e^{-\eta \|b_t - \mu_{a'}\|^2}} (b_t(s) - \mu_a(s)) \right)
\end{aligned}
$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{\tau} (2\eta \mathbb{I}\{a_t = a\}(b_t(s) - \mu_a(s)) - 2\eta\pi(a|b_t)(b_t(s) - \mu_a(s))$$
$$= \sum_{i=1}^{n} \sum_{t=1}^{\tau} 2\eta(\mathbb{I}\{a_t = a\} - \pi(a|b_t))(b_t(s) - \mu_a(s)) \,.$$

B Proofs of Propositions

B.1 Proof of Proposition 1

First, denote with $q_R^* \in \mathbb{R}^{\Delta(S)\times A}$ the optimal (belief-state) q-value function with respect to the underlying (state-space) reward function $R \in \mathbb{R}^{S\times A}$, and denote with $v^* \in \mathbb{R}^{\Delta(S)}$ the corresponding optimal value function $v_R^*(b) = \mathrm{softmax}_{a'\in A}\, q_R^*(b, a')$. Now, fix some component i of parameters θ; we wish to compute the derivative of $\log \pi(a|b)$ with respect to θ_i:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} \log \pi(a|b) &= \frac{\partial}{\partial\theta_i} \big(q_R^*(b,a) - v_R^*(b)\big) \\
&= \frac{\partial}{\partial\theta_i} \left(q_R^*(b,a) - \log \sum_{a'\in A} e^{q_R^*(b,a')}\right) \\
&= \frac{\partial}{\partial\theta_i} q_R^*(b,a) - \sum_{a'\in A} \left(\frac{e^{q_R^*(b,a')}}{\sum_{a''\in A} e^{q_R^*(b,a'')}} \cdot \frac{\partial}{\partial\theta_i} q_R^*(b,a')\right) \\
&= \frac{\partial}{\partial\theta_i} q_R^*(b,a) - \sum_{a'\in A} \pi(a'|b) \frac{\partial}{\partial\theta_i} q_R^*(b,a') \\
&= \frac{\partial}{\partial\theta_i} q_R^*(b,a) - \mathbb{E}_{a'\sim\pi(\cdot|b)} \left[\frac{\partial}{\partial\theta_i} q_R^*(b,a')\right]
\end{aligned}$$

where we make explicit here the dependence on R, but note that it is itself a parameter; that is, $R = \theta_j$ for some j. We see that this in turn requires computing the partial derivative $\partial q_R^*(b,a)/\partial\theta_i$. Let γ be some appropriate discount rate, and denote with $\rho_R \in \mathbb{R}^{\Delta(S)\times A}$ the effective (belief-state) reward $\rho_R(b,a) \doteq \sum_{s\in S} b(s)R(s,a)$ corresponding to R. Further, let

$$\begin{aligned}
&\mathbb{P}(b'|b,a) \\
&\quad = \sum_{z\in Z} \mathbb{P}(z|b,a)\mathbb{P}(b'|b,a,z) \\
&\quad = \sum_{z\in Z} \left(\sum_{s\in S}\sum_{s'\in S} b(s)T(s'|s,a)O(z|a,s')\right) \delta\left(b' - \frac{\sum_{s\in S} b(s)T(\cdot|s,a)O(z|a,\cdot)}{\sum_{s\in S}\sum_{s'\in S} b(s)T(s'|s,a)O(z|a,s')}\right)
\end{aligned}$$

denote the (belief-state) transition probabilities induced by T and O, where δ is the Dirac delta function that returns one if the belief-update (Equation 1) returns b', and returns zero otherwise. Then the partial $\partial q_R^*(b,a)/\partial\theta_i$ is given as follows:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} q_R^*(b,a) &= \frac{\partial}{\partial\theta_i} \left(\rho_R(b,a) + \gamma \int_{b'\in\Delta(S)} \mathbb{P}(b'|b,a) v_R^*(b') db'\right) \\
&= \underbrace{\frac{\partial}{\partial\theta_i}\rho_R(b,a) + \gamma \int_{b'\in\Delta(S)} v_R^*(b') \frac{\partial}{\partial\theta_i}\mathbb{P}(b'|b,a) db'}_{\rho_{R,i}(b,a)} \\
&\qquad + \gamma \int_{b'\in\Delta(S)} \mathbb{P}(b'|b,a)\mathbb{E}_{a'\sim\pi(\cdot|b')} \left[\frac{\partial}{\partial\theta_i} q_R^*(b',a')\right] db'
\end{aligned}$$

from which we observe that $\partial q_R^*(b,a)/\partial\theta_i$ is a fixed point of a certain Bellman-like operator. Specifically, fix any function $f \in \mathbb{R}^{\Delta(S)\times A}$; then $\partial q_R^*(b,a)/\partial\theta_i$ is the fixed point of the operator

$\mathcal{T}^{\pi}_{R,i} : \mathbb{R}^{\Delta(S)\times A} \to \mathbb{R}^{\Delta(S)\times A}$ defined as follows:

$$(\mathcal{T}^{\pi}_{R,i} f)(b,a) = \rho_{R,i}(b,a) + \gamma \int_{b'\in\Delta(S)} \mathbb{P}(b'|b,a) \sum_{a'} \pi(a'|b') f(b',a') db'$$

which takes the form of a "generalized" Bellman operator on q-functions for POMDPs, where for brevity here we have written $\rho_{R,i}(b,a)$ to denote the expression $\frac{\partial}{\partial\theta_i}\rho_R(b,a) + \gamma \int_{b'\in\Delta(S)} v^*_R(b') \frac{\partial}{\partial\theta_i}\mathbb{P}(b'|b,a) db'$. Mathematically, this means that a recursive procedure can in theory be defined—cf. "∇q-iteration", analogous to q-iteration; see e.g. [52]—that may converge on the gradient under appropriate conditions. Computationally, however, this also means that taking a single gradient is at least as hard as solving POMDPs in general.

Further, note that while typical POMDP solvers operate by taking advantage of the convexity property of $\rho_R(b,a)$—see e.g. [53]—here there is no such property to make use of: In general, it is *not* the case that $\rho_{R,i}(b,a)$ is convex. To see this, consider the following counterexample: Let $S \doteq \{s_-, s_+\}$, $A \doteq \{a_=\}$, $Z \doteq \{z_-, z_+\}$, $T(s_-|s_-,a_=) = T(s_+|s_+,a_=) = p = 1$, $O(z_-|a_=,s_-) = O(z_+|a_=,s_+) = 1/4$, $b_1(s_+) = 1/2$, $R(s_-,a_=) = 0$, $R(s_+,a_=) = 1/2$, and $\gamma = 1/2$. For simplicity, we will simply write b instead of $b(s_+)$. Note that:

$$\begin{aligned}
q^*_R(b,a_=) &= b\sum_{t=0}\gamma^t R(s_+,a_=) + (1-b)\sum_{t=0}\gamma^t R(s_-,a_=) = b \\
v^*_R(b) &= \log \sum_{a\in\{a_=\}} e^{q^*_R(b,a)} = \log e^{q^*_R(b,a_=)} = q^*_R(b,a_=) = b \\
\mathbb{P}(z_+|b,a_=) &= \frac{1}{4}(bp + (1-b)(1-p)) + \frac{3}{4}(b(1-p) + (1-b)p) = \frac{1}{4}b + \frac{3}{4}(1-b) \\
\mathbb{P}(z_-|b,a_=) &= \frac{1}{4}(b(1-p) + (1-b)p) + \frac{3}{4}(bp + (1-b)(1-p)) = \frac{1}{4}(1-b) + \frac{3}{4}b \\
b'|b,a_=,z_+ &= \frac{\mathbb{P}(s'=s_+,z_+|b,a_=)}{\mathbb{P}(z_+|b,a_=)} = \frac{\frac{1}{4}bp + \frac{3}{4}(1-b)(1-p)}{\mathbb{P}(z_+|b,a_=)} = \frac{\frac{1}{4}b}{\frac{1}{4}b + \frac{3}{4}(1-b)} \\
b'|b,a_=,z_- &= \frac{\mathbb{P}(s'=s_+,z_-|b,a_=)}{\mathbb{P}(z_-|b,a_=)} = \frac{\frac{1}{4}(1-b)(1-p) + \frac{3}{4}bp}{\mathbb{P}(z_-|b,a_=)} = \frac{\frac{3}{4}b}{\frac{1}{4}(1-b) + \frac{3}{4}b} \\
\mathbb{P}(b'|b,a_=) &= \begin{cases} \mathbb{P}(z_+|b,a_=) & \text{if } b' = b'|b,a_=,z_+ \\ \mathbb{P}(z_-|b,a_=) & \text{if } b' = b'|b,a_=,z_- \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Now, let the elements of θ be ordered such that p is the i-th element, and consider $\rho_{R,i}(b,a)$—evaluated at $p = 1$:

$$\begin{aligned}
\rho_{R,i}(b,a_=) &\doteq \frac{\partial}{\partial p}\rho_R(b,a_=) + \gamma \int_{b'\in\Delta(S)} v^*_R(b') \frac{\partial}{\partial p}\mathbb{P}(b'|b,a_=) db' \\
&= \frac{1}{2}\left(v^*_R(b'|b,a_=,z_+)\frac{\partial}{\partial p}\mathbb{P}(z_+|b,a_=) + v^*_R(b'|b,a_=,z_-)\frac{\partial}{\partial p}\mathbb{P}(z_-|b,a_=)\right) \\
&= \frac{1}{2}\left(\frac{\frac{1}{4}b}{\frac{1}{4}b + \frac{3}{4}(1-b)}\left(\frac{1}{4}b - \frac{1}{4}(1-b) - \frac{3}{4}b + \frac{3}{4}(1-b)\right)\right. \\
&\qquad \left. + \frac{\frac{3}{4}b}{\frac{1}{4}(1-b) + \frac{3}{4}b}\left(-\frac{1}{4}b + \frac{1}{4}(1-b) + \frac{3}{4}b - \frac{3}{4}(1-b)\right)\right)
\end{aligned}$$

Clearly $\rho_{R,i}(b,a_=)$ cannot be convex since $\rho_{R,i}(1/2,a_=) = 0$ and $\rho_{R,i}(1,a_=) = 0$ but $\rho_{R,i}(3/4,a_=) > 0$.

B.2 Proof of Proposition 2

In contrast, unlike the indirect q-value parameterization above (which by itself requires approximate solutions to optimization problems), the mean-vector parameterization of INTERPOLE maps beliefs

directly to distributions over actions. Now, the derivatives of $\log \pi(a|b)$ are given as closed-form expressions in Appendices A.1 and A.2.

In particular, note that each b_t is computed through a feed-forward structure, and therefore can easily be differentiated with respect to the unknown parameters θ through backpropagation through time: Each time step leading up to an action corresponds to a "hidden layer" in a neural network, and the initial belief corresponds to the "features" that are fed into the network; the transition and observation functions correspond to the weights between layers, the the beliefs at each time step correspond to the activations between layers, the actions themselves correspond to class labels, and the action likelihood corresponds to the loss function (see Appendices A.1 and A.2).

Finally, note that computing all of the forward-backward messages α_t and β_t in Appendix A.1 has complexity $O(n\tau S^2)$, computing all of the Jacobian matricies $\nabla_{b_t} b_{t'}$ in Appendix A.2 has complexity $O(n\tau^2 S^3)$, and computing all of the partial derivatives given in Appendix A.2 has complexity at most $O(n\tau^2 S^2 AZ)$. Hence, fully differentiating the expected log-likelihood $Q(\theta; \hat{\theta})$ with respect to the unknown parameters θ has an overall (polynomial) complexity $O(n\tau^2 S^2 \max\{S, AZ\})$.

C Experiment Particulars

C.1 Details of Decision Environments

ADNI We have filtered out visits without a CDR-SB measurement, which is almost always taken, and visits that do not occur immediately after the six-month period following the previous visit but instead occur after 12 months or later. This filtering leaves 1,626 patients with typically three consecutive visits each. For MRI outcomes, average is considered to be within half a standard deviation of the population mean. Since there are only two actions in this scenario, we have set $\eta = 1$ and relied on the distance between the two means to adjust for the stochasticity of the estimated policy—closer means being somewhat equivalent to a smaller η.

DIAG We set $T^{\text{true}}(s_-|s_-, \cdot) = T^{\text{true}}(s_+|s_+, \cdot) = 1$, meaning patients do not heal or contract the diseases as the diagnosis progresses, $O^{\text{true}}(z_-|a_=, s_+) = O^{\text{true}}(z_+|a_=, s_-) = 0.4$, meaning measurements as a test have a false-negative and false-positive rates of 40%, and $b_1^{\text{true}}(s_+) = 0.5$. Moreover, the behavior policy is given by $T = T^{\text{true}}$, $O = O^{\text{true}}$, $b_1 = b_1^{\text{true}}$, $\eta = 10$, $\mu_{a_=}(s_+) = 0.5$, and $\mu_{a_-}(s_-) = \mu_{a_+}(s_+) = 1.3$. Intuitively, doctors continue monitoring the patient until they are 90% confident in declaring a final diagnosis. In this scenario, T and η are assumed to be known. The behavior dataset is generated as 100 demonstration trajectories.

BIAS We set all parameters exactly the same way we did in DIAG with one important exception: now $O(s_-|a_=, z_+) = 0.2$ while it is still the case that $O^{\text{true}}(z_-|a_=, s_+) = 0.4$, meaning $O \neq O^{\text{true}}$ anymore. In this scenario, b_1 is also assumed to be known (in addition to T and η) to avoid any invariances between b_1 and O that we have encountered during training. The behavioral dataset is generated as 1000 demonstration trajectories.

C.2 Details of Benchmark Algorithms

R-BC We train an RNN whose inputs are the observed histories h_t and whose outputs are the predicted probabilities $\hat{\pi}(a|h_t)$ of taking action a given the observed history h_t. The network consists of an LSTM unit of size 64 and a fully-connected hidden layer of size 64. We minimize the cross-entropy loss $\mathcal{L} = -\sum_{i=1}^{n}\sum_{t=1}^{\tau}\sum_{a\in A}\mathbb{I}\{a_t = a\}\log\hat{\pi}(a|h_t)$ using Adam optimizer with learning rate 0.001 until convergence, that is when the cross-enropy loss does not improve for 100 consecutive iterations.

PO-IRL The IOHMM parameters T, O, and b_1 are initialized by sampling them uniformly at random. Then, they are estimated and fixed using conventional IOHMM methods. The reward parameter R is initialized as $\hat{R}^0(s, a) = \varepsilon_{s,a}$ where $\varepsilon_{s,a} \sim \mathcal{N}(0, 0.001^2)$. Then, it is estimated via Markov chain Monte Carlo (MCMC) sampling, during which new candidate samples are generated by adding Gaussian noise with standard deviation 0.001 to the last sample. A final estimate is formed by averaging every tenth sample among the second set of 500 samples, ignoring the first 500 samples. In order to compute optimal q-values, we have used an off-the-shelf POMDP solver available at `https://www.pomdp.org/code/index.html`.

Off. PO-IRL All parameters are initialized exactly the same way as in PO-IRL. Then, both the IOHMM parameters T, O, and b_1, and the reward parameter R are estimated jointly via MCMC sampling. When generating new candidate samples, with eqaul probabilities, we have either sampled new T, O, and b_1 from IOHMM posterior (without changing R) or obtained a new R the same way we did in PO-IRL (without changing T, O, and b_1). A final estimate is formed the same way as in PO-IRL.

PO-MB-IL The IOHMM parameters T, O, and b_1 are initialized by sampling them uniformly at random. Then, they are estimated and fixed using conventional IOHMM methods. Given the IOHMM parameters, we parameterized policies the same way we did in INTERPOLE, that is as described in (2). The policy parameters $\{\mu_a\}_{a\in A}$ are initialized as $\hat{\mu}_a^0(s) = (1/|S| + \varepsilon_{a,s})/\sum_{s'\in S}(1/|S| + \varepsilon_{a,s'})$ where $\varepsilon_{a,s'} \sim \mathcal{N}(0, 0.001^2)$. Then, they are estimated according solely to the action likelihoods in (4) using the EM algorithm. The expected log-posterior is maximized using Adam optimizer with learning rate 0.001 until convergence, that is when the expected log-posterior does not improve for 100 consecutive iterations.

INTERPOLE All parameters are initialized exactly the same way as in PO-MB-IL. Then, the IOHMM parameters T, O, and b_1, and the policy parameters $\{\mu_a\}_{a\in A}$ are estimated jointly according to both the action likelihoods and the observation likelihoods in (4). The expected log-posterior is again maximized using Adam optimizer with learning rate 0.001 until convergence.

C.3 FURTHER EXAMPLE: POST-HOC ANALYSES

Policy representations learned by INTEPOLE provide users with means to derive concrete criteria that describe observed behavior in objective terms. These criteria, in turn, enable the quantitative analyses of the behavior using conventional statistical methods. For ADNI, we have considered two such criteria: *belatedness* of individual diagnoses and *informativeness* of individual tests. Both of these criteria are relevant to the discussion of early diagnosis, which is paramount for Alzheimer's disease [51] as we have already mentioned during the illustrative examples.

Formally, we consider the final diagnoses of a patient to be belated if (i) the patient was not ordered an MRI in one of their visits despite the fact that an MRI being ordered was the most likely outcome according to the policy estimated by INTERPOLE and (ii) the patient was ordered an MRI in a later visit that led to a near-certain diagnosis with at least 90% confidence according to the underlying beliefs estimated by INTERPOLE. We consider An MRI to be uninformative if it neither (factually) caused nor could have (counterfactually) caused a significant change in the underlying belief-state of the patient, where an insignificant change is half a standard deviation less than the mean factual change in beliefs estimated by INTERPOLE.

Having defined belatedness and informativeness, one can investigate the frequency of belated diagnoses and uninformative MRIs in different cohorts of patients to see how practice varies between one cohort to another. In Table 5, we do so for six cohorts: all of the patients, patients who are over 75 years old, patients with apoE4 risk factor for dementia, patients with signs of MCI or dementia since their very first visit, female patients, and male patients. Note that increasing age, apoE4 allele, and female gender are known to be associated with increased risk of Alzheimer's disease [54–57]. For instance, we see that uninformative MRIs are much more prevalent among patients with signs of MCI or dementia since their first visit. This could potentially be because these patients are monitored much more closely than usual given their condition.

Table 5: Frequency of belated diagnoses and uninformative MRIs in various patient cohorts.

Cohort	**Frequency of belated diagnoses**	**Frequency. of uninformative MRIs**
All patients	6.52%	18.8%
Patients over 75 years old	9.29%	18.1%
Patients with apoE4 risk factor	8.75%	19.3%
Patients with signs of MCI/dementia	9.26%	26.1%
Female patients	7.19%	17.6%
Male patients	5.97%	19.8%

Alternatively, one can divide patients into cohorts based on whether they have a belated diagnoses or an uninformative MRI to see which features these criteria correlate with more. We do so in Table 6. For instance, we see that a considerable percentage of belated diagnoses are seen among male patients.

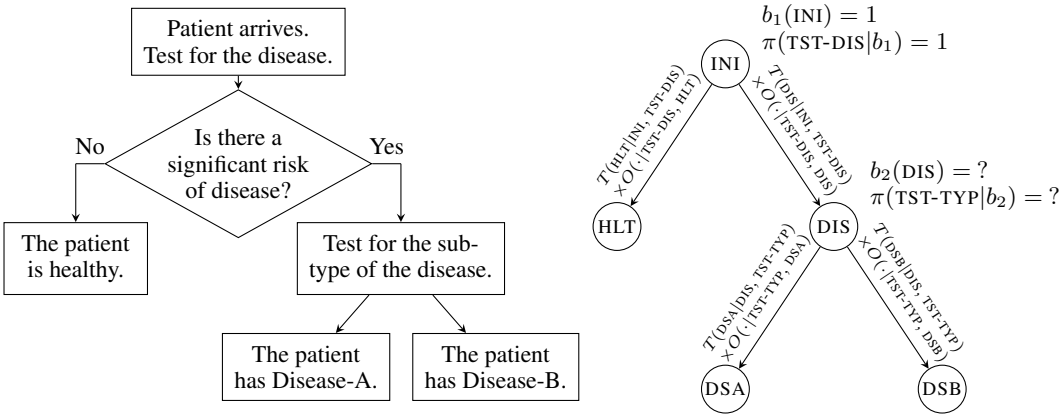
Table 6: Features of patients with belated diagnoses and uninformative MRIs.

Feature	**All patients**	**Patients with belated diagnoses**	**Patients with uninformative MRIs**
Mean age	73.9 ± 7.1	75.8 ± 7.5	73.0 ± 7.3
Freq. of apoE4	45.7%	54.2%	49.0%
Freq. of MCI/dementia signs	68.4%	95.8%	98.1%
Perc. of female patients	45.4%	39.0%	43.5%
Perc. of male patients	54.6%	61.0%	56.5%

C.4 Further Example: Decision Trees

Clinical practice guidelines are often given in the form of decision trees, which usually have vague elements that require the judgement of the practitioner [58, 59]. For example, the guideline could ask the practitioner to quantify risks, side effects, or improvements in subjective terms such as being significant, serious, or potential. Using direct policy learning, how vague elements like these are commonly resolved in practice can be learned in objective terms.

Formulating policies in terms of IOHMMs and decision boundaries is expressive enough to model decision trees. An IOHMM with deterministic observations, that is $O(z|a, s') = 1$ for some $z \in Z$ and for all $a \in A, s \in S$, essentially describes a finite-state machine, inputs of which are equivalent to the observations. Similarly, a deterministic decision tree can be defined as a finite-state machine with no looping sequence of transitions. The case where the observations are probabilistic rather than deterministic correspond to the case where the decision tree is traversed in a probabilistic way so that each path down the tree has a probability associated with it at each step of the traversal.



(a) **A Decision Tree** (b) **An Equivalent IOHMM**

Figure 5: *Two Different Descriptions of the Same Policy:* (a) in the form of a decision tree and (b) in terms of an equivalent IOHMM. For the IOHMM in (b), arrows denote possible transitions, where the probability of a transition is proportional to the quantity written above the corresponding arrow. Using direct policy learning, we can infer the risk of disease, $b_2(\text{DIS})$, and the probability of testing for the sub-type based on the risk, $\pi(\text{TST-TYP}|b_2)$, which are left vague in (a).

As a concrete example of modeling decision trees in terms of IOHMMs, consider the scenario of diagnosing a disease with two sub-types: Disease-A and Disease-B. Figure 5a depicts the policy of the doctors in the form of a decision tree. Each newly-arriving patient is first tested for the disease in a general sense without any distinction between the two sub-types it has. The patient is then tested

for a specific sub-type of the disease only if the doctors deem there is a significant risk that the patient is diseased. Note that which exact level of confidence constitutes as a significant risk is left vague in the decision tree. By modeling this scenario using our framework, we can learn: (i) how the risk is determined based on initial test results and (ii) what amount of risk is considered significant enough to require a subsequent test for the sub-type.

Let $S = \{\text{INI}, \text{HLT}, \text{DIS}, \text{DSA}, \text{DSB}\}$, where INI denotes that the patient has newly arrived, HLT denotes that the patient is healthy, DIS denotes that the patient is diseased, DSA denotes that the patient has Disease-A, and DSB denotes that the patient has Disease-B. Figure 5b depicts the state space S with all possible transitions. Note that the initial belief b_1 is such that $b_1(\text{INI}) = 1$. Let $A = \{\text{TST-DIS}, \text{TST-TYP}, \text{STP-HLT}, \text{STP-DSA}, \text{STP-DSB}\}$, where TST-DIS denotes testing for the disease, TST-TYP denotes testing for the sub-type of the disease, and the remaining actions denote stopping and diagnosing the patient with one of the terminal states, namely states HLT, DSA, and DSB.

After taking action $a_1 = \text{TST-DIS}$ and observing some initial test result $z_1 \in Z$, the risk of disease, which is the probability that the patient is diseased, can be calculated with a simple belief update:

$$\begin{aligned} b_2(\text{DIS}) &\propto \sum_{s \in S} b_1(s) T(\text{DIS}|s, \text{TST-DIS}) O(z_1|\text{TST-DIS}, \text{DIS}) \\ &= T(\text{DIS}|\text{INI}, \text{TST-DIS}) O(z_1|\text{TST-DIS}, \text{DIS}) \ . \end{aligned}$$

Moreover, we can say that the doctors are more likely to test for the sub-type of the disease as opposed to stopping and diagnosing the patient as healthy, that is $\pi_b(\text{TST-TYP}|b_2) > \pi_b(\text{STP-HLT}|b_2)$, when

$$b_2(\text{DIS}) > \frac{\mu_{\text{TST-TYP}}(\text{DIS}) + \mu_{\text{STP-HLT}}(\text{DIS})}{2}$$

assuming $\mu_{\text{TST-TYP}}(\text{DIS}) > \mu_{\text{STP-HLT}}(\text{DIS})$. Note that there are only two possible actions at the second time step: actions TST-TYP and STP-HLT.

D Details of the Clinician Surveys

Each participant was provided a short presentation explaining (1) the ADNI dataset and the decision-making problem we consider, (2) what rewards and reward functions are, (3) what beliefs and belief simplices are, and (4) how policies can be represented in terms of reward functions as well as decision boundaries. Then, they were asked two multiple-choice questions, one that is strictly about representing histories, and one that is strictly about representing policies. Importantly, the survey was conducted *blindly*—i.e. they were given no context whatsoever as pertains this paper and our proposed method. The question slides can be found in Figures 6 and 7. Essentially, each question first states a hypothesis and shows two/three representations relevant to the hypothesis stated. Then, the participant is asked which of the representations shown most readily expresses the hypothesis. Here are the full response that we have received, which includes some additional feedback:

- **Clinician 1**
  Question 1: C > B > A
  Question 2: B
  Additional Feedback: The triangle was initially more confusing than not, but the first example (100% uncertainty) was helpful. It isn't clear how the dots in the triangle are computed. Are these probabilities based on statistics? Diagram is always better than no diagram.
- **Clinician 2**
  Question 1: C > B > A
  Question 2: B
  Additional Feedback: I always prefer pictures to tables, they are much easier to understand.
- **Clinician 3**
  Question 1: C > B > A
  Question 2: B
  Additional Feedback: Of course the triangle is more concise and easier to look at. But how is the decision boundary obtained? Does the decision boundary always have to be parallel to one of the sides of the triangle?

- **Clinician 4**
  Question 1: C > B > A
  Question 2: B
  Additional Feedback: [Regarding Question 1,] representation A and B do not show any interpretation of the diagnostic test results, whereas representation C does. I think doctors are most familiar with representation B, as it more closely resembles the EHR. Although representation C is visually pleasing, I'm not sure how the scale of the sides of the triangle should be interpreted. [Regarding Question 2,] again I like the triangle, but it's hard to interpret what the scale of the sides of the triangle mean. I think option A is again what doctors are more familiar with.
- **Clinician 5**
  Question 1: C > B > A
  Question 2: A
- **Clinician 6**
  Question 1: C > B > A
  Question 2: A
- **Clinician 7**
  Question 1: C
  Question 2: B
  Additional Feedback: I thought I'd share with you my thoughts on the medical aspects in your scenario first (although I realise you didn't ask me for them). [...] The Cochrane review concludes that MRI provides low sensitivity and specificity and does not qualify it as an add on test for the early diagnosis due to dementia (Lombardi G et. al. Cochrane database 2020). The reason for MRI imaging is (according to the international guidelines) to exclude non-degenerative or surgical causes of cognitive impairment. [...] In your example the condition became apparent when the CDR-SB score at Month 24 hit 3.0 (supported by the sequence of measurements over time showing worsening CDR-SB score). I imagine the MRI was triggered by slight worsening in the CDR-SB score (to exclude an alternative diagnosis). To answer your specific questions: Q1. The representation C describes your (false) hypothesis that it was the MRI that made the diagnosis of MCI more likely/apparent the best—I really like the triangles. Q2. I really like the decision boundary.
- **Clinician 8**
  Question 1: C > B > A
  Question 2: B
- **Clinician 9**
  Question 1: C
  Question 2: B
  Additional Feedback: Q1. Representation C gives the clearest illustration of the diagnostic change following MRI. However, the representation of beliefs on a continuous spectrum around discrete cognitive states could be potentially confusing given that cognitive function is itself a continuum (and 'MCI', 'Dementia' and 'NL' are stations on a spectrum rather than discrete states). Also, while representation C is the clearest illustration, it is the representation that conveys the least actual data and it isn't clear from the visualisation exactly what each shift in 2D space represents. Also, the triangulation in 'C' draws a direct connection between NL and Dementia, implying that this is a potential alternative route for disease progression, although this is more intuitively considered as a linear progression from NL to MCI to Dementia. Q2. For me, the decision boundary representation best expresses the concept of the likelihood of ordering and MRI with the same caveats described above. Option B does best convey the likelihood of ordering an MRI, but doesn't convey the information value provided by that investigation. However, my understanding is that this is not what you are aiming to convey here.

REFERENCES

[52] M. Herman, T. Gindele, J. Wagner, F. Schmitt, and W. Burgard, "Inverse reinforcement learning with simultaneous estimation of rewards and dynamics," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, 2016, pp. 102–110.

- **Hypothesis**: This patient's condition (MCI) only became apparent **after** the first MRI was ordered.
- **Question #1**: Which of the following representations of the patient's medical history most readily expresses this hypothesis? (Please rank them from most to least accessible.)

Most recent CDR-SB:
3.0 (very mild dementia)

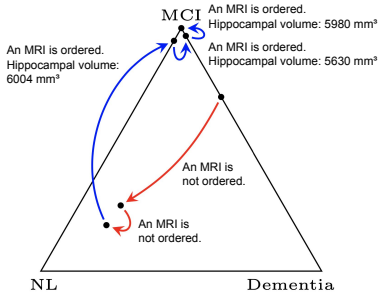
Most recent hippocampal volume:
5980 mm³ (below avg.)

Most recent ADAS11:
13.00

Most recent ADAS13:
21.00

Most recent whole brain volume:
1040240 mm³

Biomarker	Baseline	Month 6	Month 12	Month 18	Month 24
CDR-SB	1.0	1.5	2.0	1.5	3.0
CDR-SB Category	Question -able im -pairment	Question -able im -pairment	Question -able im -pairment	Question -able im -pairment	Question -able im -pairment
MRI	Not ordered	Not ordered	Ordered	Ordered	Ordered
Hippocampal Volume			6004 mm³ (below avg.)	5630 mm³ (below avg.)	5980 mm³ (below avg.)
ADAS11	12.00	11.00	10.67	11.00	13.00
ADAS13	21.00	18.00	19.67	15.00	21.00
Whole Brain Volume			1060310 mm³	1019500 mm³	1040240 mm³



Representation A: Most Recent Measurement

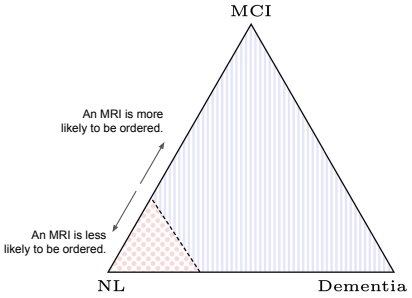
Representation B: Sequence of Measurements

Representation C: Sequence of Beliefs

Figure 6: Slide 5 out of 9, which contains the first question regarding histories.

- **Hypothesis**: The more "normal" a patient appears, the less likely an MRI is ordered.
- **Question #2**: Which of the following representations of the doctor's decision-making most readily expresses this hypothesis?

State	Reward of ordering an MRI	Reward of not ordering an MRI
NL	-0.149	-0.128
MCI	0.475	-0.115
Dementia	0.037	-0.096



Option A: Rewards

Option B: Decision Boundary

Figure 7: Slide 9 out of 9, which contains the second question regarding policies.

[53] M. Araya, O. Buffet, V. Thomas, and F. Charpillet, "A pomdp extension with belief-dependent rewards," in *Adv. Neural Inf. Process. Syst. 24*, 2010, pp. 64–72.

[54] S. Gao, H. C. Hendrie, K. S. Hall, and S. Hui, "The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta analysis," *Arch. General Psychiatry*, vol. 55, no. 9, pp. 809–815, 1998.

[55] L. J. Launer, K. Andersen, M. E. Dewey, L. Letenneur, A. Ott, L. A. Amaducci, C. Brayne, J. R. M. Copeland, J.-F. Dartigues, P. Kragh-Sorensen, A. Lobo, J. M. Martinez-Lage, T. Stijnen, and A. Hofman, "Rates and risk factors for dementia and Alzheimer's disease," *Neurology*, vol. 52, no. 1, pp. 78–78, 1999.

[56] J. Lindsay, D. Laurin, R. Verreault, R. Hebert, B. Helliwell, G. B. Hill, and I. McDowell, "Risk factors for Alzheimer's disease: a prospective analysis from the Canadian Study of Health and Aging," *Amer. J. Epidemiology*, vol. 156, no. 5, pp. 445–453, 2002.

[57] J.-H. Chen, K.-P. Lin, and Y.-C. Chen, "Risk factors for dementia," *J. Formosan Med. Assoc.*, vol. 108, no. 10, pp. 754–764.

[58] R. Chou, A. Qaseem, V. Snow, D. Casey, J. T. Cross, P. Shekelle, and D. K. Owens, "Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society," *Ann. of Internal Medicine*, vol. 147, no. 7, pp. 478–491, 2007.

[59] A. Qaseem, S. D. Fihn, P. Dallas, S. Williams, D. K. Owens, and P. Shekelle, "Management of stable ischemic heart disease: summary of a clinical practice guideline form the American College of Physicians/American College of Cardiology Foundation/American Heart Association/American Association for Thoracic Surgery/Preventive Cardiovascular Nurses Association/Society of Thoracic Surgeons," *Ann. of Internal Medicine*, vol. 157, no. 10, pp. 735–743, 2012.